Analysing Point Motion –

Spatio-Temporal Data Mining of Geospatial Lifelines

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Patrick Laube

von Böbikon AG

Promotionskomitee

Prof. Dr. Robert Weibel (Vorsitz)
Prof. Dr. Peter F. Fisher
Dr. Britta Allgöwer
Dr. Stephan Imfeld (Leitung der Dissertation)

Zürich 2005

Summary

Mobility is key to contemporary life. In a globalised world people, goods, data, and ideas move in increasing volumes at increasing speeds over increasing distances and commonly leave digital traces in space-time. Exploring the spatio-temporal process motion is an emerging research area in Geographical Information Science. Motivation

The rationale of this research is that Geographical Information Science can centrally contribute to discovering knowledge about the space-time use of individuals and groups in large volumes of tracking data. Whereas is the representation and visualisation of motion quite widespread in GIScience, approaches to actually quantitatively analyse motion are sparse. Hence, this thesis intends to contribute to the conceptual and methodological knowledge of GIScience how to analyse the tracks of moving point objects, being the most basic and commonly used conceptualisation to represent motion in geography. Thesis rationale

The thesis addresses therefore the following research questions: Can we identify and formalise a set of generic motion events, so-called *motion patterns*, that can be found in the tracks of moving point objects? How can we automatically detect such predefined motion patterns? And, finally, how can we evaluate relevance and the meaningfulness of motion patterns? Research questions

The chosen methodological approach is *geographic knowledge discovery*, an interactive and iterative process integrating a collection of methods from geography, computer science, statistics, and scientific visualisation. Its goal is the extraction of high-level information from low-level data in the context of large geographic datasets. Methods

The main contribution of this thesis consists in the conceptualisation, implementation, and evaluation of a *geographic knowledge discovery* (GKD) approach. In detail, the thesis proposed as a precondition a family of *relative motion* (REMO) patterns, with the term relative motion denoting the interrelation of motion attributes of different moving point objects over space and over time. In this thesis motion patterns are referred to as predefined formalised search templates of motion attributes such as speed, change of speed, motion azimuth, or sinuosity. Furthermore, the Constributions thesis developed a knowledge discovery process that allows conceptualisation, formal description and detection of motion patterns in the tracks of moving point objects. This required the development of a pattern description formalism as well as novel data structures and adapted pattern detection algorithms.

Significance and implications for GIScience

With its methodological approach the thesis contributes to the discussion about spatio-temporal geographical information science. The thesis points out that the integration of knowledge discovery methods within geographical information science provides a powerful means to investigate spatio-temporal phenomena, such as motion. The use of motion patterns that are intrinsically spatio-temporal, do not therefore exist in either space or time alone and, promote the need to adopt in GIScience a dynamic perception of the world. The thesis proposed a separated perception of static aggregation of objects in space and the dynamic spatio-temporal processes leading to that aggregation, for example the strict separation of the process convergence and the final static cluster as its optional outcome. The successful application of intrinsically dynamic patterns such as converging and diverging underpins that overcoming the legacy of the static geographical information systems requires not only data models and structures for dynamic data but also dynamic analysis concepts. Despite recent advances in tracking technology, the availability of motion data of large numbers of concurrently tracked individuals remains limited. With its final evaluation part, the thesis showed the potential of generating synthetic motion data to perform numerical experiments.

Outlook

The concepts and methods proposed in this thesis can be extended in various ways, especially when relaxing its rather strong basic conditions and adopting an increasingly realistic perception of real life motion processes by allowing, for instance, imperfect trajectories, fuzzy motion patterns, object-environment interactions, or by including object's semantics.

Zusammenfassung

Motivation

Mobilität ist ein wesentlicher Bestandteil des modernen Lebens. In einer globalisierten Welt bewegen sich immer mehr Menschen, Güter, Informationen und Ideen immer schneller durch Raum und Zeit. Die Flut mobiler Geräte mit Lokalisierungstechnik (GPS) produziert in zunehmendem Masse Information über die zurückgelegten Spuren ihrer Träger. Seien es nun Mobiltelefonierer in der Grosstadt, Hirsche im Schweizerischen Nationalpark oder mit Navigationssystemen ausgerüstete Fahrzeuge auf dem Strassennetz, sie alle hinterlassen ihre digitalen Spuren in Raumzeit.

Problemstellung

Folgerichtig entwickelt sich die Untersuchung raum-zeitlicher Prozesse zunehmend zu einem drängenden Forschungsschwerpunkt der Geographischen Informationswissenschaft. Diese Dissertation schöpft ihre Hauptmotivation aus der Überzeugung, dass die Geographische Informationswissenschaft wesentlich zu einem besseren Verständnis der Raumzeit-Nutzung sich bewegender Objekte beitragen kann. Während sich die Disziplin ausgiebig mit der Repräsentation und Visualisierung von Bewegung auseinander gesetzt hat, fehlen Methoden zur quantitativen Analyse von Bewegung bisher weitgehend. Die vorliegende Arbeit möchte genau diese Lücke schliessen und untersucht daher die Bewegung von Punktobjekten, der einfachsten Repräsentation von Objekten im Raum.

Forschungsfragen

Die Arbeit untersucht die folgenden Forschungsfragen: Gibt es universelle Bewegungsmuster, die sich unabhängig der Art des untersuchten Bewegungsprozesses in den Spuren von Punktobjekten niederschlagen? Wie äussern sich derartige Bewegungsmuster in den Attributen eines Bewegungspfades, d.h. in Geschwindigkeit, Beschleunigung und Bewegungsrichtung? Wie lassen sich Bewegungsmuster formell beschreiben? Wie können Bewegungsmuster in grossen Datenmengen algorithmisch aufgespürt werden? Und, wie kann die Relevanz und Aussagekraft derartiger Bewegungsmuster abgeschätzt und quantifiziert werden?

Methodik

Die Dissertation präsentiert einen Ansatz der *Geographischen Wissensgenerierung* (engl. *geographic knowledge discovery*, GKD). Diese Methodik versteht sich als spezifisch geographischer Zweig der *Wissensgenerierung aus Datenbanken* (engl. *knowledge discovery in databases*, KDD), eines interdisziplinären Methodenkat-

alogs, entstanden aus der Verschmelzung von Geographie, Informatik, Statistik und wissenschaftlicher Visualisierung. Das Ziel von GKD und KDD ist die Extraktion höherwertigen Wissens aus grossen Mengen von Rohdaten, in diesem Falle geographischen Rohdaten.

Beiträge zur Geographischen Informationswissenschaft

Der Hauptbeitrag dieser Disseration besteht in der Entwicklung, Implementation und Evaluation des *relative motion* (REMO) Ansatzes zur Untersuchung der Bewegung von Punktobjekten. Das namengebende Element des REMO Ansatzes ist eine Gruppe von Bewegungsmuster, die sich als Relationen der Bewegungsattribute verschiedener Punktobjekte ergeben. Ein Formalismus zur formalen Beschreibung dieser Bewegungsmuster bildet ein weiteres zentrales Element des Ansatzes. Eine Reihe aufeinander abgestimmter, spezieller Datenstrukturen und einfacher Mustererkennungs-Algorithmen ermöglichen schliesslich das automatische Auffinden derartiger Bewegungsmuster durch Computer. Der gesamte Analyseprozess wurde in einer eigenständigen Java-Prototypapplikation implementiert. Die erfolgreiche Anwendung der vorgestellten Methoden auf diverse Anwendungsbeispiele unterstreicht das grosse Potenzial welches in der Integration von Geographischer Informationswissenschaft und KDD steckt.

Bewegungsmuster, die ausschliesslich in Raumzeit existieren können, unterstreichen die Notwendigkeit einer dynamischen Wahrnehmung der Welt in der Geographischen Informationswissenschaft. Folgerichtig propagiert die Arbeit eine strikte Trennung des dynamischen, zwingend raum-zeitlichen Prozesses der Verdichtung (engl. *convergence*) und ihres möglichen Resultats der statischen Anhäufung (engl. *cluster*). Die Überwindung des kartographischen Vermächtnisses der statischen Schnappschuss-Weltsicht gelingt nur, wenn raum-zeitliche Datenmodelle und -strukturen und spezifisch raum-zeitliche Analysemethoden einander zuarbeiten.

Der abschliessenden Evaluationsteil illustriert das Potenzial der Generierung synthetischer Bewegungsdaten für die Evaluation eines Wissensgenerierungs-Prozesses wie etwa des REMO Ansatzes. Die Kombination von Monte Carlo Simulationen synthetischer Pfade (*constrained random walks*) und die gezielten Variation des Mustersuche-Prozesses, erlaubt die Abschätzung der Relevanz bestimmter Bewegungsmuster.

Ausblick

Mögliche Weiterentwicklungen der vorgestellten Konzepte und Methoden ergeben sich aus der Aufweichung der strengen Rahmenbedingungen, namentlich in der Behandlung unvolländiger Pfade, unscharf begrenzter (engl. *fuzzy*) Bewegungsmuster, der Miteinbezug der Bewegungsumgebung und die Berücksichtigung von Eigenschaften der bewegten Objekte.

Acknowledgements

This work was carried out during my time as a research assistant at the Department of Geography of the University of Zurich. Thus, I owe my debts first of all to the University for supporting my work. It is impossible to adequately acknowledge all the people that in many ways contributed to this thesis. Though, I will name some of them and inwardly thank the others.

I'd like to thank my chief supervisor Robert Weibel for giving me the opportunity to carry out this research in the Geographic Information Systems Group. He acted as an always interested and accessible mentor, supported my unorthodox employment mode during the first half of my PhD time and disencumbered me for the crucial second half.

I am indebted to my advisor Stephan Imfeld for enduring assistance and time for discussions. Certainly without him, I would not have completed this thesis. I probably even would not have started it.

A special thank has to be addressed to Peter Fisher, Professor for Geographic Information Science at City University, London, who found among his many other duties the time to be a member of my PhD-committee.

I owe a huge debt of thanks to Ross Purves who's door always was open for me. With him I found a Java teacher, a demanding and experienced reviewer and a smart counterpart for countless discussions about motion and other interesting things of life. What started as a funny 'Java-German-tandem' evolved into friendship. Thank you.

I would particularly like to thank Britta Allgöwer. At the very beginning some years ago it was due to her insatiable enthusiasm for the various application problems at the fruitful interface between geography, biology, and GIScience that I took an interest in motion.

The collaboration with Marc van Kreveld from the University of Utrecht was a stroke of luck. The brainstorming session at tea time in Dagstuhl Castle gave rise to some nice ideas about explicit spatial motion patterns and resulted in our contribution to SDH 2004. I got to know Marc as a fast thinking mind and hard playing sportsmen: He returned the manuscripts almost as fast and sharpened as the table-tennis balls in the Dagstuhl gym.

In addition, I have to thank some other people at the Department of Geography that in many ways contributed to the success of this work. Sabine Timpf for many interesting discussions, careful reviews, and for her discrete mentoring. Ruedi Haller for helpful discussions, for fun days doing field research in the Engadina, and for grounding my sometimes aloof ideas providing invaluable insights in the 'real wildlife GIS' of the Swiss National Park. Dani Isenegger for being a stalwart companion during the many years of our studies and some great hikes. My room mate Alistair Edwardes for being a pleasant and humorous companion. Nikos Koutsias for fruitful discussions about the basics in statistics. The IT-people Othmar Wigger and Dani Wirz for running the computers and keeping them happy.

This work is about analysing existing data. Thus, it would not have been possible without many people generously letting me use their data - in many cases collected or produced with great efforts. Heiri Leuthold and Michael Hermann provided the wonderful data about the moving districts in the ideological space that was for a very long time my sample data during the implementation stage. Our discussions about the motion of abstract objects in abstract spaces were very illuminating. Sachiko Iwase and Hideo Saito from Department of Information and Computer Science, Keio University, Japan, provided the soccer data that opened up the field of soccer scene analysis. Susan Rupp from Texas Tech University, Lubbock, TX, patiently managed the formalities to use the Los Alamos elk data and was furthermore an interesting partner for transatlantic chats. A special thank goes to the the nameless researchers of the *Porcupine Caribou Project* and the *Starkey Project* for providing their invaluable tracking data on the web.

Furthermore I am indebted to Martin Steinmann. Besides willingly giving very valuable advice in graphic design issues, he always was an interesting partner for talkin' jazz and a perpetual source for 'musical software'.

I owe a huge debt of thanks to my parents who always supported my studies and encouraged me for the next steps.

And lastly, I must thank the many deer, elk, soccer players, and moving districts that inspired me with their sometimes enigmatic lifelines and taught me a lot about the nature of motion and the motion in nature.

Contents

List of Figures

List of Abbreviations

Part I

Synopsis

Chapter 1

Introduction

"Give me space and motion and I will give you a world."

René Descartes, in: E. T. Bell, Men of Mathematics,
Touchstone (reissued edition), 1986.

1.1 Motivation

Mobility is a key to contemporary human life and underpins the phenomenon of space-time compression (May and Thrift, 2001). In a globalised world people, goods, data, and ideas move in increasing volumes at increasing speeds over increasing distances. In human affairs this has highlighted the inadequacy of analyses which side-step the issue of mobility, but even in the animal domain (far less accelerated by technologies) the critical place of motion and mobile agents in gaining ecological knowledge is enthusiastically acknowledged (Hulbert, 2001). Appropriately, the challenge of exploring motion is also a key emerging research area in Geographical Information Science (GIScience). This thesis intends to contribute to GIScience's analytical power to investigate motion at an individual and aggregate level.

1.1.1 Analysing motion – imperative and opportunity

A challenging imperative

In the closing weeks of 2002, the first 'new' disease of the 21st century - SARS - emerged in southern China. Only a few days later this severe respiratory infection had leapt to Hong Kong, Vietnam, Singapore, Canada and Germany and beyond. From November 2002 to July 2003, a cumulative total of more than 8000 probable SARS cases with more than 750 deaths were reported in 26 countries (Pearson et al., 2003). This diffusion was driven by people on the move. These individuals were somehow interrelated in their travel, used spatial networks, leaving their marks in space-time. What can we learn from this event to be prepared for the emergence of a new infectious disease? We know from the work of Cliff and Haggett (2004) and Cliff et al. (1993) in the Pacific and elsewhere that a detailed analysis of the movements of a disease carrier can tell us much about the historic diffusion mechanisms of a transmissible disease. In the modern world the scale of the problem and the complexity of the movements means that understanding and predicting motion phenomena is a *vital and challenging imperative*.

A promising opportunity

The first Global Positioning System (GPS) sensors heralded in an age where it would be commonplace for devices to be able to record a near continuous stream of time-stamped locations reflecting the whereabouts of the sensor or the agent carrying it. The ubiquity of such devices is reflected by the fact that in the summer of 2004 the Japanese Telecommunications Council declared that all mobile phones introduced in Japan after 2007 should have self-locating functionality. This technological condition is intended to allow locating of users in the case of an emergency call. Many other nations have already legislated for similar requirements and the warehousing of locational records for a period of years as an anti-terrorism measure. In addition, telecommunications companies have already moved to secure their rights to these data.

Whether for good or evil, location aware devices are becoming ubiquitous and will increase our capability to collect spatio-temporal movement data by orders or magnitude. Sophisticated combinations of GPS receivers, handheld computers and mobile phones will soon constantly create large data volumes covering the motion of individuals in space over time (Miller, 2003; Mountain and Raper, 2000). Notwithstanding legitimate ethical caveats (Dobson and Fisher, 2003) such detailed information about tracked vehicles, people, or animals can enable geographers, traffic planners, sociologists, and wildlife biologists to investigate individual and group behaviour. The knowledge gained has huge applied potential for areas such as geomarketing, wildlife pest control, fundamental biological research, or dynamic traffic control systems to reduce congestion, – providing the science is there to harness the data. Understanding and predicting motion is a *promising opportunity* in both the academic and applied spheres.

Geographic knowledge discovery

To seize these opportunities to exploit motion data requires the development of "spatio-temporal data mining and exploratory visualisation techniques that can handle the massive, noisy space-time-attribute data" (Miller, 2003, p. 6). "In addition to 'traditional' database manipulation, analysis, and visualisation tasks, geographical information systems (GIS) and GIS users also now need to filter through vast amounts of data to find patterns and associations" (Peuquet, 2002, p. 6).

A number of international research groups have been looking at aspects of this problem (Imfeld, 2000; Miller and Han, 2001; Forer et al., 2004; Kraak and Koussoulakou, 2004), and have unanimously reflected the fact that studying the new forms of motion data requires a shift in perspective on how motion is analysed. The *geographic knowledge discovery* (GDK) approach presented in this thesis proposes such a shift in perspective, integrating methods from GIScience, database research, and information visualisation.

1.1.2 The fetish of the static

"It's about time"

The support for spatio-temporal data is indispensable for almost every urgent geographic field, e.g. climatic change and sustainability, globalisation and population issues, or urban studies and health applications. Modern geography and geosciences request the modelling of complex and dynamic geo-phenomena as well as the spatio-temporal exploration of [...] geographic data (Raper, 2000).Ten years ago Donna J. Peuquet pointed out in her seminal article that "it's about time". She claimed that the "greater promise of spatio-temporal Geographical Information Systems (GIS) resides ultimately in their capacity to examine causal relationships and their effects in any of four modes of inquiry: exploration, explanation, prediction, and planning" (Peuquet, 1994, p. 443).

Time cuts across GIScience

Defining the field of GIScience Michael Goodchild asserted around the same time that modelling time-dependent geographical data is unique to GIS and thus has a great potential (Goodchild, 1992). More than ten years later Goodchild repeated this issue at the first centennial of the Association of American Geographers, pointing out that a rapid progress in representing time in GIS, and in the development of methods for the analysis of spatio-temporal data is needed (Goodchild, 2004b). David Mark agreed that "time is an integral part of GIScience research and one which cuts across most other GIScience topics" (Mark, 2003, p. 12). The ongoing importance of "space and space/time analysis and modelling" as a GIScience research issue is furthermore underlined by it's listing on the University Consortium for Geographic Information Science (UCGIS) Research Agenda as a long-term research challenge (University Consortium for Geographic Information Science, 1996), – just like spatial ontologies, scale, and uncertainty.

Fetish of the static

In spite of the community's consensus on the important role of time for geographic information handling, progress is slow and spatio-temporal systems are exceptions rather than the rule. After almost forty years of development GIS are still weak in handling spatio-temporal data. In his keynote talk at the 2nd International Conference on GIScience 2002 in Boulder, CO, Jonathan Raper listed the exploration of new representations of the dimensions of space and time as one of the five challenging dimensions of GIScience. He concluded by recommending that the "fetish of the static" be abandoned (Raper, 2002).

Cartography's view of the world

Emerging from cartography's view of the world, the data models of most nowadays commercial GIS are static. Representing change is limited to enumerating temporal delimited snapshots (Berry, 1964; Chrisman, 1998; Couclelis, 1999). "The representation of both space and time in digital databases is still problematic and functional space-time systems have not yet gone beyond the limited prototype stage" (Peuquet, 2001, pg. 11). One reason for this shortcoming may be that the representation of phenomena in time as well as space is significantly more complex and more difficult than their representation in space alone, because time and space exhibit important differences in their properties and in their referential bases for potential queries (Peuquet, 1994).

Individual-based perspective in geography

The growing interest in the possibilities of tracking data gives rise to another vital development in GIScience: A shift from a place-based perspective to an individual-based perspective in geography. Traditional place-based methods are no longer valid in a world where people and activities are becoming disconnected from location. For this reason Harvey J. Miller postulates the development of a "rigourous, formal representational theory of the dynamic spatial objects of interest in time geography and activity theory" (Miller, 2003). Especially at finer geographical scales analysis "changes the emphasis from a concern for understanding the structural arrangement of objects [i.e. place-based perspective] to ways in which those objects move to position themselves in time and space [i.e. individual-based perspective] (Batty et al., 2003, p. 673). An individual-based perspective in geography investigates the interactions between objects rather than the static structure of aggregates. Focussing on interacting people, animals, or any other geographic entity suggests that *mobility* as well as location becomes important (Batty et al., 2003).

1.1.3 GIScience' weaknesses in analysing motion

Concluding the motivation section the following outlines open problems in GIScience related to spatio-temporal analysis of motion.

Snapshot view of the world

Nowadays commercial out-of-the-box GIS are based on cartography's static snapshot view of the world. Static conceptual data models and data structures can't cope with the dynamic world.

Spatio-temporal is special

Not only spatial is special (Anselin, 1990; O'Sullivan and Unwin, 2003), but spatio-temporal is also special. Collapsing detailed trajectories in summarising spatial statistics conceals potentially important events or inter-object relations. The conventional spatial analysis toolbox, developed mainly to understand the static arrangement of entities in space is not well suited to analysing moving and dynamically interacting spatial entities.

Motion

Due to past technological limits and the ideological legacy of static cartography the analysis of individual motion has a limited tradition in GIScience. Recent advances in tracking and telecommunication technologies and the emergent interest in an individual-based, object oriented view of geography have highlighted the analysis of motion on the agenda of the GIScience community.

Tracking data volumes

Tracking data are typically voluminous, incomplete (gaps), and imprecise. Such tracking data easily overwhelms traditional geographic analysis methods designed to tease scarce information from small and homogeneous datasets (Miller and Han, 2001). GIScience lacks robust analysis methods to derive process knowledge from spatio-temporal information.

1.2 Thesis rationale

The basic rationale of this research is that GIScience can centrally contribute to discovering knowledge about the space-time use of individuals and groups in the emerging volumes of tracking data. Therefore this thesis intends to contribute to the conceptual and methodological knowledge of how to handle motion in the mostly static world of GIS. This section presents first a set of research questions that gave the rationale for this thesis. Second, a set of research objectives structures the problem solving process of this work.

1.2.1 Research questions

Narrowing the focus

In a dynamic world, motion is just one type of change to be expressed by spatial objects. Life and state complete the triad of change. *Motion* refers to changes in the position or geometric form of an object over time; *life* refers to changes in the temporal identity of an object; and finally *state* refers to changes in the semantics or non-spatial attributes of the object (Miller, 2003). This thesis focuses on motion, because it is most interesting and challenging from a conceptual GIScience perspective. Life and state raise more technical and implementation based questions to be typically answered by computer scientists and database specialists (Abraham and Roddick, 1999). Narrowing further the focus, this thesis concentrates on points, the 'ultimate spatio-temporal primitives' (Raper, 2002).

This thesis intends to contribute to GIScience' theory to analyse the motion of moving point objects (MPOs). The thesis therefore addresses the following research questions.

$\Longrightarrow$ *Are there generic motion patterns, i.e. equal or at least similar patterns that can be found in the tracks of MPOs modelling various motion phenomena?*

$\Longrightarrow$ *How can we describe, formalise, and thus compare delimited motion events of individuals and groups of individuals?*

$\Longrightarrow$ *How do we best model MPOs in order to detect motion patterns?*

$\Longrightarrow$ *Can we automatically detect motion patterns in typically voluminous tracking data of groups of MPOs?*

$\Longrightarrow$ *How can we identify interrelated or interacting individuals (subgroups) in larger groups of MPOs?*

$\Longrightarrow$ *Can we identify motion patterns that are intrinsically dynamic, that is do not express any pattern in either space or time alone?*

$\Longrightarrow$ *How can we evaluate the interestingness of motion patterns?*

1.2.2 Research objectives

The main objective of this thesis is to develop a framework allowing geographic knowledge discovery in the trajectories of MPOs. Four research objectives encapsulate the above research questions. Each objective gives rationale for one of four successive work packages, forming the natural stages of this thesis.

Objective 1: *This thesis shall identify, characterise, and categorise a set of basic motion patterns in the lifelines of groups of MPOs, with motion patterns being a predefined formalised search template of motion attributes such as speed, change of speed, or motion azimuth.*

Objective 2: *This thesis shall develop a knowledge discovery approach integrating techniques to formally describe motion patterns and algorithms allowing detection of these patterns in case study data.*

Objective 3: *This thesis shall develop means to describe and algorithmically detect forms of temporally co-ordinated* *and spatially directed group motion patterns in trajectories of MPOs, such as flocking, converging or diverging.*

Objective 4: *This thesis shall develop statistical measures and adopt Monte Carlo techniques to quantify the relevance and interestingness of the proposed motion patterns, to distinguish random noise from information in the knowledge discovery process.*

1.3 Structure of the thesis

This thesis reports on research successively published in peer-reviewed conference proceedings and scientific journals. The research done for each work package has been published in an individual publication directed to a selected audience. The four publications are:

Laube, P. and Imfeld, S. (2002). Analyzing relative motion within groups of trackable moving point objects. In Egenhofer, M. J. and Mark, D. M., editors, *Geographic Information Science*, volume 2478 of *Lecture Notes in Computer Science*, pages 132–144. Springer, Berlin-Heidelberg, DE.

Laube, P., Imfeld, S., and Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668.

Laube, P., Van Kreveld, M. and Imfeld, S. (2004). Finding REMO - detecting relative motion patterns in geospatial lifelines. In Fisher, P. F., editor, *Developments in Spatial Data Handling*, Proceedings of the 11th International Symposium on Spatial Data Handling, pages 201–215. Springer, Berlin-Heidelberg, DE.

Laube, P. and Purves, R.S (submitted 2005). Evaluating motion pattern techniques in spatio-temporal data. *Computers, Environment and Urban Systems*, Submitted April 2005.

Thesis in two parts:
Synopsis and Publications

This volume has two parts. The first part (synopsis) integrates the individual perspectives of the particular publications in a scientific summary and embeds their individual contributions in a set of introductory and concluding chapters. The second part (publications) presents the publications in exactly the content and format they appeared in the original sources. These publications have thus not been cast into 'monograph-like' chapters.

Thematic order of publications

For the final compilation of this thesis the publications are brought in a thematic order to create a logical sequence for the reader. This order does not completely correspond to the initial publication dates. Links from the first part to the original publications in the second part are established as conventional references (e.g. Laube et al., 2004).

The structure of the synopsis is as follows:

Chapter 1: **Introduction** The thesis starts with an introductory chapter giving the motivation and developing the rational for this research (p. 3).

Chapter 2: **State of the Art** Constructed as a state of the art the chapter reviews different aspects of handling motion with respect to analysis in the disciplines geography, biology and computer science and their specialised fields cartography, GIScience, database management, and behavioural ecology (p. 11).

Chapter 3: **Methods & Results** The third chapter presents an executive summary of the four publications. This chapter gives a short overview of the rationales, methods, results, and contributions of each publication. It does not replace the publications, it only gives the reader a structured access to the research presented in the original papers attached in the end of this thesis. Concepts, definitions, and terms may have changed during some years of research. This chapter identifies such ambiguities guiding the reader (p. 61).

Chapter 4: **Discussion** The fourth chapter discusses the results of this thesis' research from a holistic and integrated perspective, – providing a synthesis. This evaluation chapter compares the proposed analytical framework with the latest approaches from the literature recalling the research questions, performs a SWOT analysis of the proposed approach, sketches its potential applications and limitations, and finally lists the problems that remained open (p. 71).

Chapter 5: **Conclusions** The concluding fifth chapter is concerned with the importance of the thesis to the development of the discipline of GIScience. The chapter lists the contributions of the thesis as a whole, its insights and gives an outlook (p. 91).

Chapter 2

State of the art

> *"In science the object of the exercise is not to find a theory that will, or is likely to, be deemed true for ever; it is to find the best theory available now, and if possible to improve on all available theories."*
>
> David Deutsch, 1998, p. 64

Various disciplines within science and technology are increasingly interested in analysing motion. The literature presented in this state of the art review originates in the disciplines of geography (GIScience, cartography, urban studies), biology (ecology, wildlife biology), computer science (database research, information visualisation, data mining and knowledge discovery), and statistics. The overall goal of this literature review is to report efforts to approach motion in order to understand motion in all these fields. Therefore, this chapter covers eight interrelated tasks, all relevant for the analysis of motion. These tasks are:

1. Capturing motion
2. Quantifying motion
3. Modelling motion
4. Formalising motion
5. Querying motion
6. Visualising motion
7. Analysing motion
8. Simulating motion

This literature review mainly focuses on approaches designed for point objects, but with some links to more complex moving entities such as lines or polygons where sensible.

Thematically organised literature review

The state of the art has a thematic structure in order to facilitate the access to the different aspects of motion. Since most scientific approaches touch more than just one of the tasks listed above, the reader may find many sources referenced more than once. However, every section focuses on those aspects of the cited research relevant for each specific task. But before reviewing the literature, the three central terms of this thesis – **Analysis**, **Motion**, and **Points** – need to be clarified.

Analysis in GIScience generates geographic knowledge

Analysis. In logic analysis refers to the discovery of general principles underlying concrete phenomena. In this thesis analysis is performed in the general context of *geographic information science* (GIScience), i.e. that branch of information science that covers the fundamental principles underlying the design, testing, and use of geographic information technologies (Goodchild, 2004a). Analysis implies a procedure that generates new knowledge, in this case geographic knowledge about the geographic phenomenon of motion. According to Golledge (2002) *geographic knowledge* aim to understand and analysing accumulated facts to produce new information and knowledge that are not directly observed during data gathering (Golledge, 2002). Thus, analysis in the notion used in this thesis must go beyond descriptive statistics or database queries.

Change, life, and motion

Motion. In GIScience the definition of motion comes often along with the conceptualisation of *change*. As Frank (2001) points out change comes in two forms: "Change of the objects of interest and change in the position or geometric form of these. For the first we use the heading *life* of objects: objects may appear and disappear (for example, a forest or a residential zone), two objects may merge (for example, two parcels or two towns), and objects may split. For the second we use the heading *motion*: objects may move or may appear to move, with or without changing their form at the same time" (Frank, 2001,p. 22)

Eulerian vs. Langrangian motion

Motion can be perceived from the Eulerian or the Lagrangian perspective, respectively (Turchin, 1998). The *Eulerian* view considers changes as they occur at a fixed point in space. Static points are characterised by numbers or fluxes of things moving by. Geographical diffusion processes or vector field representations of a glacier's movement are typical Eulerian abstractions of motion. The *Langrangian* view considers changes which occur following a moving object's trajectory. Here, an individual's movement can be described by motion azimuth, velocity, and acceleration. Recording GPS tracks of people, animals or vehicles is a typical example for the Langrangian concept of motion. This thesis mainly focuses on Langrangian motion.

Points. In Euclidean geometry a point is defined as a dimensionless (0-D) entity that has a location in space but no extent and no orientation. The location is normally specified by a set of $(x, y(, z))$ coordinates.

Entity with location without extent

In geography a point is one of the three fundamental abstractions of spatial objects, lines and polygons being the two others. A geographical point describes an object whose location, but not extent, is relevant. A simple point entity implies that the spatial extent of the object is limited to a location at the level of resolution of the abstraction. Thus, a town can be represented by a point at a continental level of resolution but as a polygon entity at a regional level (Burrough and McDonnell, 1998).

Moving point objects (MPOs)

In a dynamic view of the world a point object needs not be static, it may move around and evolve into a *moving point object* (MPO). At any time t an MPO has a location specified by a tuple of $(x, y(, z), t)$ coordinates. In this thesis the MPO is the prime object of interest, irrespective its real world counterpart.

2.1 Capturing motion

Technology and science has produced an amazing diversity of approaches to track moving objects. The tracking the motion of objects has been a tedious and expensive task over decades. However, presently it is evolving into an ubiquitous by-product in today's communication and IT-society.

2.1.1 Tracking animals

At a broad enough temporal scale, both animals and plants are mobile life forms. Thus, biology investigates the mobility of life forms, mainly addressing two different issues, first the spatial redistribution of populations and second the motion of individuals. Section 2.3.1 will return to this basic dichotomy in the modelling of motion.

Mark-recapture

In *mark-recapture* studies animals are captured, somehow marked, released and hopefully recaptured after a certain time period. See Turchin (1998) for a methodological overview. At best, mark-recapture studies provide information about the overall displacement of an organism. At worst, such studies reveal which individuals disappear, without giving any clue how to separate death from emigration.

Individual paths

Thus, whenever possible, behavioural ecologists prefer capturing individual paths. For an introductory overview on path tracking methods see also the second chapter of Turchin (1998). Following individuals, mapping their paths, and digitising their tracks is extremely labour and time intensive (e.g. Wentz et al., 2003). Working with small animals such as insects may allow laboratory based work and thus a reduction in fieldwork costs (Morales and Ellner, 2002). Of course, attaching collars on animals my influence their body conditions and thus their behaviour (Tuyttens et al., 2002).

VHF telemetry

The use of VHF (very high frequency) radio telemetry is widespread when tracking animals that are strong enough to carry a radio transmitter collar. The handheld VHF receivers facilitate following or (re-)finding previously collared animals. Some intensive long-term monitoring projects even use a VHF telemetry system with stationary VHF antennas to provide an automatic continuous coverage of even large groups of collared moving animals (e.g. Brillinger et al., 2001b,a, 2004).

GPS tracking

GPS tracking technology revolutionised tracking in wildlife biology (Hulbert, 2001; Sibbald et al., 2001). Once equipped with a GPS collar moving animals report their positions in real time or in a batch mode. The following references give an impression of the diversity of recent biological research done in tracking animals with GPS. McGrady et al. (2003) tracked Steller's Sea Eagles using the ARGOS system (www.argosinc.com) to investigate annual migration and to estimate the size of winter ranges and summering grounds. Matthiopoulos et al. (2004) estimated the space use of grey seals in order to balance the seal's needs with the requirements of the marine exploitation. Franke et al. (2004) tracked caribou with a high temporal granularity ($\delta t = 15$ minutes) in order to compare monitored behaviour with modelled behaviour. As an example of applied animal behaviour science Ganskopp (2001) tracked two cows in order to evaluate the effectiveness of salt and water spot manipulations for managing cattle distribution in arid-land pasture.

Incomplete trajectories

Even with the increasing possibilities of the recent tracking technologies, recording the trajectory of an animals motion is difficult and sometimes cumbersome. First, animals may be too small to carry a collar (von Hühnerbein et al., 2000) or just be hard to follow (Byers, 2001). Collared animals may disappear, i.e. die, get shot, or simply leave the area of investigation. Furthermore they may loose their collars or the collar may show a malfunction. Finally animals may also move in rugged and mountainous terrain, across dense vegetation challenging VHF as well as GPS tracking technology (Haller et al., 2001; Wentz et al., 2003).

Oversampling and undersampling

The selection of the appropriate sampling rate of fixes along a path poses the problem of avoiding oversampling and undersampling (Turchin, 1998). Undersampling causes information loss since the path is sampled at a resolution that is too coarse. It can be avoided by collecting the data at the highest resolution feasible. In an oversampled path a new data point does not add new information to the information contained in the previous point. Oversampling causes noise and implies non-existent autocorrelation in movement. Several approaches allow resampling oversampled paths in order to represent the object's motion at the magnitude that the analysis is focussing on. See Turchin (1998, p. 130) for resampling methods of oversampled paths.

2.1.2 Tracking vehicles

Vehicles tracking systems

Another type of trajectories are captured when it comes to tracking moving vehicles in transportation geography or military applications. In most cases the taxi cabs, tanks, vessels, or air-planes are not actively tracked but rather constantly report their location in an operational mode. With very little cost a set of fundamental measures to quantify motion at almost arbitrarily fine granularity can be derived from such raw sensor data. Smyth (2001) lists *position*, *fix time*, *velocity*, *heading*, *acceleration*, and *orientation change* as measures easy to log using an automative navigation system. Even if tracking is not a scientific task here, the production of geospatial lifelines is an interesting technological by-product. Such operational tracking data normally covers the motion of vehicles with tuples of $(x, y, (z,)t)$ in regular time steps.

GIS-T

The application of GIS in transportation is referred to as GIS-T (Goodchild, 2000). It has to be noted that the understanding of motion is different in many transportation applications: The objects do not move in a featureless space but operate on constrained networks, e.g. road networks (Miller and Wu, 2000). That may influence the capture of motion. For instance, while moving on a straight highway no constant location updating is required. The section on modelling MPOs will return to this issue (section 2.3.1)

2.1.3 Tracking people

Travel diaries

In contrast to animals and vessels, people can be *asked* when they were where. Thus, questionnaires are a straight forward way to capture people's itineraries (Forer et al., 2004). The resulting activity-travel diary data may consist of event chains, indicating when an individual was stationary at some location, and when they moved from one place to another (Frihida et al., 2004a). Such an event chain is normally structured by motion events (e.g. commuting, shopping, leisure) and may contain considerable periods of staying somewhere (e.g. at home, at work).

More sophisticated ways of obtaining travel activity schedules include computer-based questionnaires (Doherty and Miller, 2000). Technological advancements in location aware devices also revolutionised the field of travel activity scheduling. More and more sophisticated combinations of GPS receivers, mobile phones and handheld computers allow constant tracking of individuals in an almost twenty-four-seven mode (Mountain and Raper, 2000; Dykes and Mountain, 2003).

Soccer scene analysis

One other way to capture motion in delimited spaces is to derive an individual's location from motion picture data from surveillance or television cameras, e.g. during a football game (Iwase and Saito, 2002, 2003).

2.1.4 Tracking spatial change

Change

Capturing motion as a crucial part of spatio-temporal change has a long tradition in geography. Mapping and re-mapping the position and form of spatial entities is at the core of many geographic disciplines, such as geomorphology, landscape ecology, or remote sensing. Repeatedly observing or measuring, mapping and digitising the locations of spatial entities is a very rudimentary and expensive, but nevertheless practical way of capturing motion. Even though there has been work done to explicitly track moving elements in the landscape (e.g. Kääb, 2002), such research focuses on the static representation of the summarised positional change between, for instance, two points in time t_1 and t_2. Thus, time is rather a delimiter of snapshots that represent mainly the *state* and *life* aspects of change rather than the dimension of motion as defined in this thesis.

Capturing motion: Lessons learned for analysing the motion of points

- Notwithstanding advances in tracking technologies, some data capture procedures intrinsically generate incomplete and inhomogeneous trajectories. Any motion analysis technique must cope with that.
- The fields investigating motion are very diverse, and so are the generated motion data. For the design of analysis methods this causes first of all a data integration problem.
- The sampling rate gives a hard constraint on the range at which analysis can be performed.

2.2 Quantifying motion

The motion of an MPO is grounded in its trajectory. Both, the current motion as well as its trace can be quantified using measurable motion parameters. Such measures are, for instance, speed and direction in the case of current motion, total trajectory length, step length, or sinuosity in the case of describing a trajectory. Quantification is an important precondition to compare either the motion of individuals or between different kinds of MPOs. Using more sophisticated descriptive measures provides additional means of motion analysis (see section 2.7.1).

2.2.1 The physical background is simple...

Relations among force, matter, and motion

Mechanics deals with the relations between force, matter, and motion. The part of mechanics concerned with describing motion is called *kinematics.*

Average velocity

Given an MPO moving in the xy-plane along a curved path from Point P to point Q (Figure 2.1), the displacement of the MPO is the change in the position vector $\mathbf{r}$, given the x- and y-component of $\Delta\mathbf{r}$ as Δx and Δy, and Δt referring to the duration of the described motion.

$$\Delta x = x_2 - x_1 \tag{2.1}$$

$$\Delta y = y_2 - y_1 \tag{2.2}$$

$$\Delta t = t_2 - t_1 \tag{2.3}$$

In kinematics the *average velocity* v_{av} is defined to be the vector quantity equal to the displacement divided by the time interval (Sears et al., 1987). Thus

$$v_{av} = \frac{\Delta\mathbf{r}}{\Delta t} \tag{2.4}$$

The average velocity is a vector quantity having the same direction as $\Delta\mathbf{r}$ and a magnitude equal to the magnitude of $\Delta\mathbf{r}$ divided by Δt. The magnitude of $\Delta\mathbf{r}$ is the straight line distance from P to Q, regardless the actual shape of the path in between P and Q.

Instantaneous velocity

The velocity of an MPO at a specific point in the path or at a specific instant of time is called the *instantaneous velocity.* It is defined in magnitude and direction as the limit approached by the average velocity when Point P is taken closer and closer to point Q.

$$v = \lim_{\Delta t \to 0} \frac{\Delta\mathbf{r}}{\Delta t} = \frac{d\mathbf{r}}{dt} \tag{2.5}$$

Speed

The meaning of the term *speed* is twofold. “It may mean the magnitude of the instantaneous velocity. For example, when two cars travel at 50 km/h one north and one south, both have the speed of 50 km/h. In a different sense, referring to an average quantity, the speed of a body is the total length of path, divided by the elapsed time. Thus if a car travels 90 km in 3 hours, its

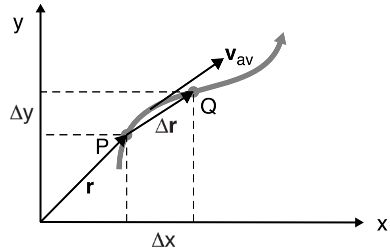


Figure 2.1: Average velocity v_{av}. The displacement $\Delta\mathbf{r}$ between P and Q has the components Δx and Δy. The average velocity Δv has the same direction than $\Delta\mathbf{r}$. Source: Sears et al. (1987, p. 50).

average speed is 30 km/h, even if the trip starts and ends at the same point. In the latter case the average velocity would be zero, because the total displacement is zero" (Sears et al., 1987, p. 28). However, this thesis always refers to *speed* as a scalar quantity, the magnitude of velocity. Thus, the thesis addresses speed and motion azimuth separately.

Average acceleration

The *average acceleration* a_{av} of an MPO moving from P to Q is defined as the vector change in velocity, $\Delta\mathbf{v}$, divided by the elapsed time Δt (Sears et al., 1987).

$$\mathbf{a}_{av} = \frac{\Delta\mathbf{v}}{\Delta t} \tag{2.6}$$

Instantaneous acceleration

The *instantaneous acceleration a* of an MPO refers in analogy to instantaneous velocity to its acceleration at some point of its path at some instant of time. It is defined in magnitude and direction as the limit approached by the average acceleration when point Q approaches point R and $\Delta\mathbf{v}$ and Δt both approach 0.

$$\mathbf{a} = \lim_{\Delta t \to 0} \frac{\Delta\mathbf{v}}{\Delta t} = \frac{d\mathbf{v}}{dt} \tag{2.7}$$

In a strict scientific sense, acceleration has a direction. Again, this thesis focussed on the magnitude of acceleration as a scalar quantity and investigates the motion azimuth separately.

Motion azimuth

The direction of a line in space is called many things: bearing, course, heading, flightline, or azimuth (Robinson et al., 1995). This thesis focusses on the azimuth of a line used in surveying, as the angle between the northwards oriented tangent on the meridian of a point P on the rotation ellipsoid and the tangent on a line through P. The azimuth is the angle of the motion compass direction measured from the North in clockwise direction. Thus, 0° is North, 90° East, 180° South, and 270° West. This compass metaphor is also applicable for non geographic spaces such as abstract attribute spaces.

Sinuosity

Motion *sinuosity* denominates whether a path is straight or wiggling around in many curves. Dutton (1999) proposes the use of a local line sinuosity. This sinuosity is the ratio of the distance measured along the trajectory to the length of the trend line connecting its end points. This sinuosity adopts a value of 1 for straight lines with collinear fixes and increases to ∞ for infinitely wiggling trajectories. Wentz et al. (2003) define sinuosity exactly the other way around as the ratio between the straight-line between the end points and the distance measured along a track. Thus, this sinuosity value varies between 0 and 1, where 1 represents a straight line.

2.2.2 ...but the devil is in the details of implementation

Path, move, and step

Behavioural biologists have spent decades of field work capturing and quantifying the *motion paths* of species such as butterflies, ants, birds, or elk. In comparison to mark-recapture approaches, recording exact paths is more laborious but nevertheless preferred if possible. Even though animals trace continuous paths the characteristics of their motion are normally translated to a discrete form suitable for analysis, usually with a series of straight lines (Figure 2.2a). Turchin (1998) defines the concepts of path, move, and step to quantify motion. A *path* is the complete spatio-temporal record of a followed organism, from the beginning to the end of observation. A path consists of *moves*, defined as the displacements between two consecutive stopping points, for example stop-over flowers of a butterfly searching for oviposition[1] (Figure 2.2a). *Steps* are introduced to discretise paths of organisms that move continuously such as migrating elk (Figure 2.2b). Steps are defined as displacements during typically regular time intervals. Once the path is appropriately discretised several motion quantifiers (speed, acceleration, azimuth, sinuosity), or summary path descriptive measures (average move length, turning frequency, turning angle distribution) can be computed.

Setting the observational interval

The crux of deriving motion properties from the trajectory of an MPO lies in the interplay of sampling rate and observational interval. For all motion quantifiers at least two consecutive fixes on the trajectory are needed to compute the distance $\Delta\mathbf{r}$ and Δt. The width of the observational interval in relation with the availability of fixes on the underlying trajectory heavily influence

[1] Oviposition: egg-laying.

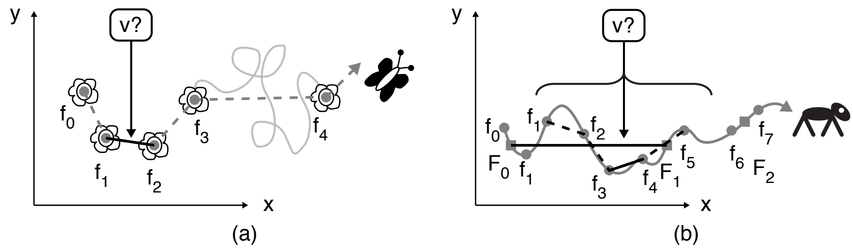


Figure 2.2: Two different kinds of trajectories: (a) In the butterfly-mode with observed consecutive stopping points, speed v at query time can be computed dividing the *move* length (from Fix f_1 to f_2) by the corresponding time, well knowing that the MPO might have travelled a much longer track (e.g. illustrated for f_3 to f_4). (b) The fixes in the rather continuous track in the elk-mode are artificially introduced by discretisation. Speed can here be computed dividing the *step* length by the corresponding time. Note that different results will emerge depending on having a fine (f_3 to f_4) or a coarse sampling rate (F_0 to F_1). Using a longer interval (track from f_1 to f_5) instead of just considering the anterior and the posterior fix results in yet another speed.

the derived motion quantifiers (Figure 2.2). The availability of fixes is first dependent on the data capture procedure (see section 2.1), and second on the conceptual data model of motion that is used (see the following section 2.3).

Underestimating speed

One has furthermore to be well aware that the fixes provide only an approximation of the actual movement, assuming that the known fixes are connected with a linear lifeline thread, capturing the likely space-time locations at which a MPO may have been moving continuously from Point A to Point B (Hornsby and Egenhofer, 2002), from fix f_3 to f_4 in Figure 2.2a respectively. Hence, the speed indication derived using a straight line tends to produce speeds which are too low compared with the actual speed of the MPO.

Pitfalls of circular statistics

Motion azimuth is a circular quantity. Thus, computing the average motion azimuth over a specified temporal interval selecting a set of fixes may be cumbersome when ranging over North. For instance, Cain (1989) found that out of 22 surveyed ecological field studies investigating motion, in 17 studies turning angles were not analysed correctly! See Mardia and Jupp (2000) for an excellent introduction to directional statistics.

Sampling rate and sinuosity

It is furthermore very evident from Figure 2.2, that the interplay of sampling rate and observational interval has also a huge influence on a repeated computation of sinuosity. Considering the whole trajectory plotted in Figure 2.2 the sinuosity gained from

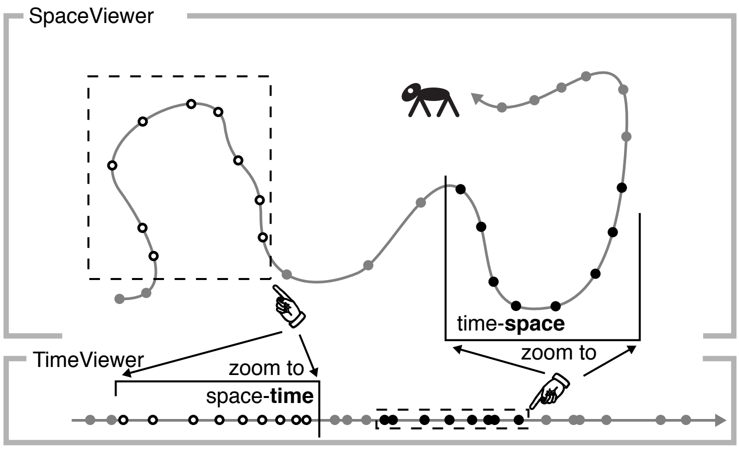


Figure 2.3: A *time-space* is a spatial range defined by setting a temporal limit. By contrast, a *space-time* is a temporal range defined by setting spatial limits. Source: Own illustration after a concept by Dykes and Mountain (2003).

the fine sampling rate f_x is low, and conversely adopts high values using the coarse sampling rate F_x.

Time-space and space-time

Most motion quantifiers can furthermore be computed for larger trajectory subsets. Dykes and Mountain (2003) derive absolute speed, direction and sinuosity for space-times and time-spaces (Figure 2.3). A *time-space* indicates where an MPO is during a particular period (spatial range defined by a temporal limit). In contrast, a *space-time* indicates when an MPO is in a particular space (temporal range defined by a spatial limit).

Quantifying motion: Lessons learned for analysing the motion of points

- The fix sampling rate and observational interval, be it given by the data capture method or superimposed in deriving the motion quantifier, may influence the resulting motion quantifiers.

2.3 Modelling motion

Conceptual data models vs. data structures

This section reviews possible ways of conceptualising and representing motion. *Conceptual data models* (section 2.3.1) are abstractions which incorporate only those properties thought to be relevant to an application – a human conceptualisation of reality. By contrast, *data structures* (section 2.3.2) are a representation of the conceptual data model reflecting the way data will be stored

Abstract vs. discrete models

(Peuquet and Marble, 1990). In the *database management systems* (DBMS) community *abstract models* refer to the conceptual data model, and *discrete models* refer to data structures implementable in databases (Erwig et al., 1999).

Retrospective versus prospective perspective

Modelling motion can adopt two fundamentally different perspectives: a retrospective and a prospective one. With *retrospective* applications the motion process is over and thus all the fixes of the moving objects are known in advance. This perspective is typical for analytical applications such as data mining of animal observations. The *prospective* perspective is adopted when the motion process has to be mapped live in an on-line mode. Typical applications are digital battlefields and traffic control systems. Data models and structures for the latter case must allow querying in the future and thus unknown or uncertain states of the database (see section 2.3.1).

Event driven vs. observation-based systems

Another critical issue for modelling motion concerns the methods for data capture. Some tracking systems are *event driven* system, where the system automatically detects and reacts on status changes, for example in the speed or motion direction of an MPO. Such systems normally feature irregularly sampled data. On the other hand *observation-based systems* capture data in an ordered sequence at regular intervals (Moreira et al., 1999).

These different perspectives require different data models and data structures. The following paragraphs give a structured overview of different ways of modelling the motion of point objects.

2.3.1 Conceptual data models

What, where, – and when?

Modelling a moving object is more complex than modelling a static spatial object. Modelling motion requires a conceptualisation of space *and* time, or to be more precise the conceptualisation needs to address both in an integrated way as *space-time* or *time-space* (Raper and Livingstone, 1995; Massey, 1999). Motion requires not only the descriptors *what* and *where* but also *when.*

Different types of time

Frank (1998) provides a useful overview of the different types of conceptualisations of time. He identifies three different subdivisions to establish a taxonomy of time models. The first model focuses on the conceptualisation of the considered events (abstract time points or intervals between two events). The second results from the interpretation of the considered process (events arriving on linear time or arriving in a cyclicly repetitive way). The third considers the scale of measurement (ordinal time scale or interval scale).

Database times

In the context of databases different dimensions of time must be distinguished (Moreira et al., 1999). The time when a fact is true

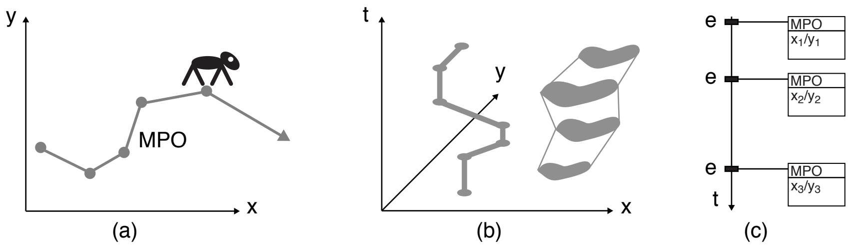


Figure 2.4: Conceptual data models of motion: (a) Entity primary: Consecutive fixes (x, y) of MPO entities lie side by side in a 2-D space. (b) Space-time primary: MPOs produce shapes in 3-D space (x, y, t). (c) Event primary: Update events on a timeline change the MPOs' properties.

in reality is referred to as the *valid time*. The valid time originates typically from the observation data. The *transaction time* of a fact refers to the time when a fact is current in the database and may be retrieved. Other authors term these two temporal dimensions *real time* or *event time* (valid time), and *system time* or *database time* (transaction time), respectively (Worboys, 1998). Databases that handle both valid and transaction time are called *bitemporal* databases (Pasquale et al., 2003).

Temporal is special

The additional dimension time features some significantly different properties than are known from spatial dimensions. Time can be modelled as continuous and progressive, or as cyclical. The partition of time can be regular or irregular. A regular partition model is best suited in cases of automatic fixing at given times. Irregular partitioning allows representation of motion when fixes depend on occasional observations. Furthermore the exact moment when to record a location and the extent of an object may be observation system driven (e.g. GPS availability) or event driven (e.g. a parcel has a new owner).

Irrespective of the conceptualisation of time as a dimension, this thesis proposes adopting three different perspectives on modelling motion.

- Entity primary
- Space-time primary
- Event primary

Entity primary. The static 2-D space is populated by moving entities or moving objects, respectively. These entities are described as having attributes and a sequence of fixes, that is observations of (x, y) associated to a time t (Figure 2.4a). All subsequently introduced entity primary approaches focus on point entities, the underlying principles could, however, also be applied for polyline objects.

Geospatial lifelines

The most basic entity primary conceptualisation of motion is the geospatial lifeline of a point object. Mark (1998, p. 12) defines a *geospatial lifeline* as a "continuous set of positions occupied in space over some time period. Geospatial lifeline data consists of discrete space-time observations of a geospatial lifeline, describing an individual's location in geographic space at regular or irregular intervals". As shown in Figure 2.2 these observation provide only approximations of the actual movement which is potentially much more complex than just a straight line.

Moving object databases (MOD)

Traditional DBMS assume that data is constant until it is explicitly changed. Such databases are not suitable to handle continuously changing positions of MPOs in real time. By contrast, *moving objects databases* (MOD) represent information about moving objects and their location. For example, a typical MOD task would be representing the current locations of a taxi-cab or airplane fleet. Such an MOD allows queries like "retrieve all airplanes that will come within 30 miles of the airport in the next 10 minutes". MOD applications arise in transportation systems and in the military context of the digital battlefield (Wolfson et al., 1998b).

MOST data model

Special MOD data models have been developed to handle moving objects (Moreira et al., 1999). The Moving Objects Spatio Temporal (MOST) data model proposed by Sistla et al. (1997) and Wolfson et al. (1998b) illustrates the idea of an MOD data model. It has two main aspects: First MOST features *dynamic attributes*, i.e. attributes that change over time according to some given function, until it is explicitly updated. Second MOST is accompanied with a *future temporal logic* (FTL) (Sistla et al., 1997; Trajcevski et al., 2004) so that the user can query future states of database values (see section 2.5).

Shapes in space-time

Space-time primary. In space-time primary models multiple geolayers represent the changing location and extent of a moving object in a stack of time-slice fields. Space-time primary motion models indicate where and when a moving object occupied which spot in the space-time, spanning a two-dimensional space and modelling time in a similar way as a third, additional spatial dimension (see Figure 2.4b). The motion of an object thus produces a continuous shape in space-time. Moving points create forms in the space-time volume that look like erect, distorted poles, the motion and deformation of lines and areas create erect walls and pillar-shaped volumes.

Space-time aquarium

The most prominent space-time primary data model is the *space-time aquarium* of Time Geography initiated by Hägerstrand and the Lund School (Hägerstrand, 1970). The main idea of Time Geography is that a moving object generates a timeline over space and through time. This timeline is normally visualised using a two-dimensional space and adopting a third dimension of time. In this framework a stationary object becomes a vertical line, a moving object an inclined and deformed line, reflecting motion speed and direction. Adopting certain constraints about an object's motions, typically maximum speed, the space-time aquarium allows modelling of an object's potential locations, the timeline becomes a space-time prism. Since the 1970s there have been many attempts to introduce the space time aquarium in GISystems mainly to perform activity and accessibility analysis (Miller, 1991; Forer, 1998; Erwig et al., 1999).

Data types *mpoint* and *mregion*

The CHOROCHRONOS project, a European research project to promote research on *spatio-temporal database management systems* (STDBMS), proposes the spatio-temporal data types *mpoint* and *mregion* (Erwig et al., 1999; Gueting et al., 2003; Koubarakis and Sellis, 2003). As with the space-time aquarium, this approach views moving points and moving regions as 3-D (2-D space plus time) or higher dimensional entities whose structure and behaviour is captured by modelling them as abstract data types to be integrated in relational, object-oriented or other DBMS data models. "A value of type *mpoint* describing a position as a function of time can be represented as a curve in 3-D space (x, y, t)" (Erwig et al., 1999, p. 273) (Figure 2.5a)

Events

Event primary. The organisational basis of event primary data models is a perpetual progressive timeline, arranging events in a relative order. *Events* are defined as instants or intervals in time when objects are changed, for example if they change their location in the sense of moving point objects, or change their shape or attributes (Wachowicz and Healey, 1994). Event primary modelling has mainly received attention in the database community. From a database perspective a motion process is a sequence of delimited change events, each requiring an update operation in the database. Every update records simply the new spatial attributes of the moving object, its location and extent, respectively.

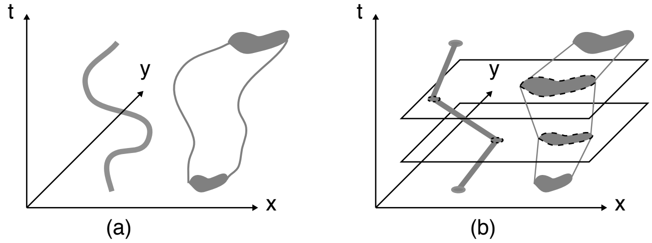


Figure 2.5: CHOROCHRONOS data types and structures: (a) The abstract data types *mpoint* and *mregion*, and (b) the corresponding sliced, hence discrete, data structures *upoint* and *uregion* introduced in section 2.3.2. Source: Own illustration after Erwig et al. (1999).

Event-based SpatioTemporal Data Model (ESTDM)

The Event-based Spatiotemporal Data Model (ESTDM) introduced by Peuquet and Duan (1995) illustrates the event primary perspective. ESTDM is a time-based approach that temporally orders changes to locations. ESTDM stores specific changes associated with times in an event list C_i. The specific stored temporal location t_i is called a *time-stamp*. It is assumed that each event list and associated changes relate to a single thematic domain (e.g. land-use). The ESTDM has not explicitly been designed to model motion. However, the relocation of discrete motion steps can easily be handled with this data model.

Activity-based travel behaviour (ABTB)

Focusing explicitly on human activity or travel behaviour requires a more diary or itinerary like representation of MPOs in a database (Forer and Simmons, 2000; Miller, 2003; Forer et al., 2004). A special form of event-primary data models are used for a better understanding of activity-based travel behaviour (ABTB). Some recent publications present research that investigates diaries or itineraries such as logs of individuals in order to understand the people's activities and their movement behaviour, respectively (Frihida et al., 2004b,a). In their data model valid time in a database is associated with events (trips or activities) that happened in reality. The start time and the end time delimits the time span boundaries (i.e. the interval) of an event. Events are either activities (e.g. work, home, shopping, entertainment, etc.) or trips (moving from A to B, e.g. walking, by bus, car, etc.). A so-called *time-path* consisting of a sequence of move (trip) events and activities records an individual's daily behaviour. According to this notion the track of an MPO can be represented as a move (trip) event chain consisting solely of move events.

2.3.2 Data structures

When modelling space alone, entities are often modelled using vector data structures and fields using raster data structures. However, there is no such clear mapping for modelling spatio-temporal phenomena, especially not for modelling motion. The data structures presented in the following paragraphs cannot easily be linked to one of the above introduced categories of object-based, space-time-based, or event-based data models.

Sliced representation of moving objects

As Erwig et al. (1999, p. 281) state very concisely, "abstract models are simple, but only discrete models can be implemented." Abstract models allow one to make definitions of spatio-temporal entities in terms of infinite sets, without worrying whether finite representations of these exist. It is thus very simple and straightforward to view a moving point as a continuous curve in the space-time aquarium. But when it comes to the implementation only finite or reasonably small sets are usually stored and manipulated in computers. The above curve, for instance, would be discretised into a set of connected straight segments of curve functions (Figure 2.5b).

More generally speaking, the CHOROCHRONOS project introduced the approach of a *sliced representation of moving objects*. In this approach, a moving object is represented as a set of 'slices', so-called *temporal units*. A temporal unit is a maximal interval of time where values taken by an instance can be described by a simple function (e.g. a linear function connecting two consecutive fixes instead of a curve). In this sense the abstract model of a continuous line in 3-D space modelling an *mpoint* (see page 25) is converted in the spatial unit type *upoint* (Figure 2.5b), consisting of a set of 3-D segments (Gueting et al., 2003).

Vector data structures. It is very simple to establish a vector data structure for discretely moving points: just record their fixes in a flat spreadsheet featuring tuples of x, y, z, t, a where a stands for attributes.

Column vector $\boldsymbol{r}$

In ecology fixes are often recorded in a simple column vector from tracking MPOs in the field. It consists of a collection of the X, Y coordinates for every object. Altogether the locations are structured in a column vector giving the k-th time of the m-th MPO.

$$\mathbf{r} = \{X_m(t_{mk}), Y_m(t_{mk})\}' \tag{2.8}$$

This data structure is often used to capture the fixes of tracked animals in behavioural ecology (Turchin, 1998; Brillinger et al., 2001b).

Moving polylines

Data structures for moving polylines are much more complex and are not discussed in detail in this thesis. However, see for example Guibas (2004) for a data structure for collision detection of two moving convex polygons and Pang and Shi (2002) for an adaptation of the Voronoi model to maintain the topological relationships between moving polygons.

Kinetic data structures (KDS)

In the field of (dynamic) computational geometry the algorithmic description of motion has also produced some data structures. These data structures are designed to describe a geometric configuration of a set of n MPOs rather than their actual motion. The objective of such *Kinetic data structures (KDS)* is the maintenance of configuration functions such as 'convex hull' or 'closest pair' under continuous motion of the given objects (Basch et al., 1997; Guibas, 1998). "A KDS for a geometric attribute is a collection of simple geometric relations that certifies the combinatorial structure of the attribute, as well as a set of rules for repairing the attribute and its certifying relation when one fails" (Guibas, 2004, p. 2). That means for the convex hull example: MPOs are allowed to move around freely until one inner point breaks the hull and thereby initiates an update.

From voxels to taxels

Raster data structures. Forer (1998) and Huisman and Forer (1998) propose a 3-D array data structure to handle the Lund School space-time aquarium. Each array cell, normally called *voxel* in a 3-D space, represents a space-time instance or span. Here the array cells are called *taxel* to emphasise the specific nature of the time dimension. Moving in space-time, an MPO occupies taxels. Given a known lifeline, the taxel represents presence of the MPO. Given, in contrast, a space-time prism of potential presence of the MPO, then the taxels represent the bits of space-time, where the MPO could be. Thus, a known lifeline may then be a series of adjacent taxels, while a space-time prism may be a cone-shaped volume of taxels.

ESTDM data structure

The data structure of the Event-based SpatioTemporal Data Model (ESTDM) proposed by Peuquet and Duan (1995) consists of a header, a base map and an event list. The base map defines the initial state of the world at t_0 in a complete raster snapshot image. The event list contains every time-stamp, pointers to the changing event components and a pair of pointers, *prev* and *next*, that point to the previous and next event in the list, respectively.

Ecological modelling

Many ecological models simulating the motion of species adopt a raster view of the world. Simple raster data structures are used to represent homogeneous and heterogeneous habitats populated with a set of moving agents representing animals or plants. According to a set of rules the agents relocate at each step from their current cell to another cell in the raster (e.g. Carter and Finn, 1999; Morales, 2002). Some authors even use a hexagonal grid (Beecham and Farnsworth, 1998). A lifeline can then just be described by a list containing the coordinates of the successively occupied cells.

2.3.3 Modelling incomplete and uncertain motion data

The description of moving points is often incomplete and uncertain. On the one hand remote tracking technologies suffer from periods of missing data. The satellite coverage of a GPS tracked animal may temporally be low, the receiver in the collar may show a malfunction, or the collar may simply fall off. On the other hand the exact location between known fixes is always uncertain and thus hard to estimate and query. Hence, data models and data structures are the first place to handle incompleteness and uncertainty in the analytical process.

Error ellipse

Capturing the movement of an MPO by sampling its position using a GPS receiver at regular time intervals introduces uncertainty about the objects position in-between known fixes (Figure 2.6a). Pfoser and Jensen (1999) propose a model for this uncertainty based on the sampling rate and the maximum speed of the MPO. Similarly to the concept of the space-time prism they limit the possibilities of where the MPO could have been. An error ellipse enclosing two consecutive positions P_1 and P_2 as its foci captures the points' positional uncertainty (Figure 2.6b). The size and shape of the error ellipse depends on the sampling rate and the object's speed.

Uncertainty threshold r

A simpler model is that proposed by Trajcevski et al. (2004). They associate a constant uncertainty threshold r with every line segment of a trajectory. Seen in 3-D space, a horizontal uncertainty area of radius r moves along the segments of the trajectories. Thus, the uncertain positions of an MPO are represented as a series of cylindrical bodies enclosing the possible motion curves. This concept bears similarities to the ε-band concept used to model error bands of static 2-D data in GIS (Burrough and McDonnell, 1998).

Variability functions

Moreira et al. (1999) present a very similar model also decomposing the trajectory of moving points into sections. Between the known fixes so called *variablity functions* store information about the track, and ellipses are used for possible trajectories between variability functions.

Lifeline beads

In their article about modelling motion over multiple granu-

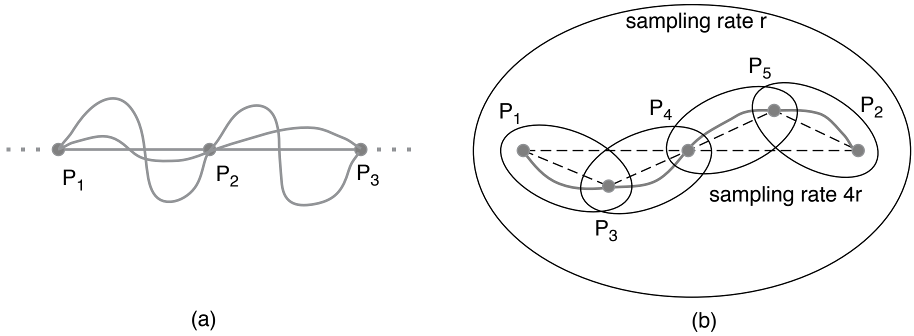


Figure 2.6: (a) Possible trajectories of an MPO between known fixes P_1, P_2, and P_3 (Pfoser and Jensen, 1999, p. 117). (b) At sampling rate r the large error ellipse limits the possibilities where the MPO could have been travelling from P_1 to P_2. Increasing the sampling rate to $4r$ means that the additional fixes P_3 to P_5 better approximate the movement, and the error introduced by sampling decreases, represented with the small error ellipses. Source: Pfoser and Jensen (1999, p. 120).

larities Hornsby and Egenhofer (2002) present an approach that integrates the entity primary and the space-time primary perspective. They start off by modelling motion as geospatial lifelines. Shifting the lifeline from a planar 2-D view to a 3-D space-time-aquarium the lifeline transforms to a *lifeline thread*, referring to a linear approximation of an ordered sequence of space-time samples capturing the likely points at which an MPO may have been moving from A to B. Modelling the set of all possible locations for an MPO given A and B and a maximum speed is based on a set of geometric constraints that describe the intersection of two inverted half cones that form a *lifeline bead*. Refining granularity means adding fixes to the lifeline. The inclusion of additional fixes refines the geometry such that the bead transforms to a sequence of smaller beads, called a *lifeline necklace*.

Missing data

Imperfect satellite coverage may also lead to missing data and fragmentary tracks. Wentz et al. (2003) compare two methods to create continuous tracks from fragmentary data. The linear weighted distance approach simply generates new points lying on a straight line between the known fixes. The more sophisticated constrained random walk approach generates more realistic circuitous paths using a combination of Monte Carlo methods and the space-time prism concept (see section 2.8).

L.uncertainty

Wolfson et al. (1998b) extend their MOST model to represent moving objects with uncertain positions using the additional sub-attribute $L.uncertainty$. Since the uncertainty of the real position increases after every update, the MPO commits to send an update whenever the deviation reaches the given bound $L.uncertainty$. That means that the uncertainty of an MPO's location (L) has at maximum the value of $L.uncertainty$.

Modelling motion: Lessons learned for analysing the motion of points

- "Abstract data models are simple, but only discrete models can be implemented" (p. 281 Erwig et al., 1999).
- The elegant space-time primary data models such as the space-time aquarium are hard to implement. They have hardly entered practical applications.
- Event-based models are primarily designed to track sporadic changes of space, and are thus less suited for objects constantly moving.

2.4 Formalising motion

Capture the essential features

In computer science a formalisation is the attempt to capture the essential features of a real-world phenomenon (in the case of this thesis 'motion') in a formal language. Such a formalism can subsequently be used to communicate and interact with an information system or a database. A formal language allows expressing the temporal relations inherent to a motion process, such as *starts at, is after*, or *overlaps*. A formal description of motion is furthermore needed if one wants to retrieve specific motion events from the information system or database, e.g. by formulating a query (see section 2.5). Finally, such a formalism helps in the design of databases suitable for motion data (Koubarakis and Sellis, 2003).

Describing change

A formal description of motion often comes along with formal descriptions of change. In their survey of spatio-temporal databases Abraham and Roddick (1999) give a brief introduction to mathematically describing spatio-temporal information. Their overview focuses mainly on tracking changes to objects and ignores motion processes. Similarly Hornsby and Egenhofer (2000) present a change description language with respect to states of existence and non-existence for objects. Their approach is designed to track the identity change and changes in the regional structure of primarily cadastral objects and thus is not suited for formalising MPOs.

Spatio-temporal predicates

In analogy to the seminal *9-intersection model* of spatial predicates by Egenhofer and Franzosa (1991), *spatio-temporal predicates* describe developments of relationships between objects changing their position and shape. Spatio-temporal predicates are used to query STDBMS. Erwig and Schneider (2003, p. 194) and Guet-

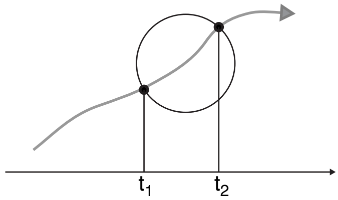


Figure 2.7: Visual specification of the *Cross* predicate. The plane (grey lifeline) crosses the shape of a storm (black circle). Source: Erwig and Schneider (2003, p. 188).

ing et al. (2003, p. 161) define the basic ontological predicates for moving points: *disjoint, meet, inside*. These predicates can for instance be used to describe the process of an aircraft crossing a storm (Figure 2.7). A query that builds on this formalism could ask for MPOs that were first *disjoint* from the evolving region storm, then *meet* to be temporarily *inside*, and finally *meet* to be *disjoint* again in the end (expression 2.9).

$$Cross := disjoint > meet > inside > meet > disjoint \tag{2.9}$$

Constraints

Grumbach et al. (2001, 2003) introduce a framework to model spatio-temporal data as infinite sets in rational space. A polygon in the plane is seen as the infinite set of points in $\mathbb{Q}^2$ inside its frontier. The attribute values of a *trajectory* form an infinite set of triples (x, y, t) which is represented in a finite way by linear constraints (Figure 2.10). The following trajectory describes the subarea where the ski tourist Pirmin is skiing from morning ($t = 9$) to noon ($t = 12$):

$$9 \leq t \leq 12 \wedge x - y \geq 0 \wedge y \geq 3 \wedge y \leq 7 \wedge x \leq 11 \tag{2.10}$$

This notion allows one to represent an individual's activity by aggregating trajectories such as expression (2.10) to a relation *People* as follows:

People

Name	Category	Activity	Trajectory
Pirmin	Tourist	Sleeping	$0 \leq t \leq 8 \wedge x = 3 \wedge y = 6$
Pirmin	Tourist	Eating	$8 \leq t \leq 9 \wedge x = 3 \wedge y = 6$
Pirmin	Tourist	Skiing	$9 \leq t \leq 12 \wedge x - y \geq 0 \wedge y \geq 3 \wedge y \leq 7 \wedge x \leq 11$

Taxonomy for the evolution of point groups

Thériault et al. (1999) give a spatio-temporal taxonomy for the representation of spatial set behaviours. In order to analyse travel behaviour in metropolitan areas they propose a formal taxonomy for the description of the evolution of entity sets

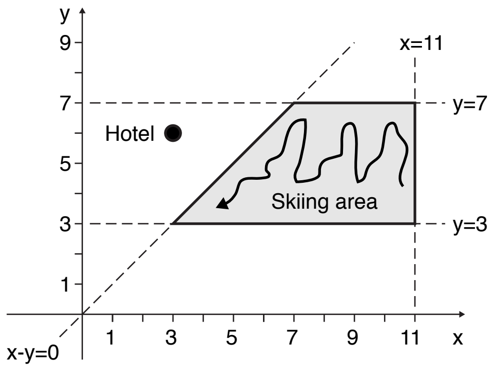


Figure 2.8: Simplified constraint view of a skitourist's world. The grey polygon represents Pirmin's motion activity *Skiing* formalised in expression 2.10. Source: Gueting et al. (2003, p. 180).

in space and statistical pattern description measures. In their taxonomy they present different kinds of changes related to a *set of geographical entities (SGE)*. Apart from some changes related to the appearance and disappearance of individuals, there are the two classes *changes related to movement of individuals* (spread/concentration, deformation of SGE, restructuring of SGE) and *changes related to movement of entire SGE* (translation, rotation). The application presented along with the taxonomy focussed on the analysis of changing SGE arrangements in an origin-destination survey of the Quebec Urban Commmunity's public transportation corporation (STCUQ).

Formalising motion: Lessons learned for analysing the motion of points

- In GIScience formalisms are often used to describe change of state (e.g. attributes of parcels), but hardly ever to describe change of position, that is motion.
- In database research formalisms are widely used to design and query databases managing motion data.

2.5 Querying motion

Queries: efficient access to MPOs

In spatio-temporal database applications the current as well as the past and anticipated future positions and extents of objects are of interest (Gueting et al., 2000). For MPOs this general task can be narrowed down to changing positions. These aspects are captured answering various queries based on the locations, trajectories, and topology, and exploring patterns in the motion of objects (Agarwal et al., 2002).

A *query* in a spatio-temporal database "is a predicate over the database history (rather than a predicate over a single database state, as in traditional databases). The *answer* to a query [...] consists of the set of instantiations of the variables that satisfy the predicate" (Sistla et al., 1997, p. 3).

Thus, in short, queries allow efficient access to MPOs. This section illustrates a set of approaches to query the motion of objects first adopting a retrospective and second adopting a prospective view integrating uncertainty about future states.

2.5.1 Querying moving objects in databases

The database community has developed terms for different kinds of queries on motion (Theodoridis and Papadias, 1995; Agarwal et al., 2002; Raptopoulou et al., 2003):

- *Location-based queries.* These queries include range queries and proximity queries.
  - The term range represents a subdivision of the space, normally a rectangle. A typical example for a *range query* is: "Given a rectangle R and a time t, report all MPOs that will be inside R at t".
  - *Nearest-neighbour queries* are typical *proximity queries*: "Given a point P and an interval $[t_1, t_2]$, report the MPO that is nearest to P in that interval $[t_1, t_2]$.
- *Continuous queries.* These queries allow a moving location of the query range. For example: "Keep track of all cars within x kilometres around car B".
- *Trajectory-based queries.* These queries involve the topology of the trajectories and derived informations such as velocity or motion azimuth of the MPOs. For example: "Which members of a set of MPOs are heading North?" – "Report all objects whose speed doubles in the next t Minutes".

The following paragraphs give some illustrations of queries, corresponding to data models and data structures listed in section 2.3.

Querying using spatio-temporal data types

The spatio-temporal data types introduced by Erwig et al. (1999); Gueting et al. (2003); Koubarakis and Sellis (2003) along with some operations for these data types provide a powerful framework for querying MPOs. For example, consider the operation

mdistance.

$$\underline{mpoint} \times \underline{mpoint} \rightarrow \underline{mreal}\ \mathbf{mdistance} \tag{2.11}$$

The operation **mdistance** computes the distance between the two moving points at all times and hence returns a real number changing over time, a type called <u>*mreal*</u> (for *moving real*). This allows one together with the straightforward operation **minvalue** to answer queries like: "Find all pairs of planes that during their flight came closer to each other than 500 metres".

SELECT A.id, B.id (2.12)
FROM flights A, flights B
WHERE A.id < > B.id AND
minvalue(**mdistance**(A.route, B.route)) < 500

Erwig et al. (1999) call this in analogy to a 'spatial join' an instance of a *spatio-temporal join*.

Constraints

As another offspring of the CHOROCHRONOS project, the approach of Grumbach et al. (2003) allows querying of spatio-temporal information based on constraints. In this approach locations are given by constraints that delimit the subarea where the MPO could possibly be. The constraint data model supports declarative query languages for spatio-temporal queries. Motion data represented by generalised relations can be queried using first-order logic (relational calculus) or relational algebra (Kanellakis et al., 1995). Returning to the example of the ski tourist Pirmin a query could be: "Where is Pirmin between 10 and 12?". In relational calculus, this query could be expressed as:

$$\begin{aligned} &name, x, y : (\exists t)(People(name, t, x, y) \\ &\qquad \wedge name = Pirmin \wedge 10 \leq t \leq 12) \end{aligned} \tag{2.13}$$

The answer to query (2.13) is the following relation:

Name	Place
Pirmin	$x - y \geq 0 \wedge y \geq 3 \wedge y \leq 7 \wedge x \leq 11$

The answer comes from the third tuple of the *People* relation. The first two tuples do not match the selection because the temporal condition is not met.

Query-by-trace

Based on the spatio-temporal predicates introduced in section 2.4 Erwig and Schneider (2003) propose a visual notation that is able to describe scenarios of changing MPO relations and is implemented in a graphical interface called *Query-by-trace* (QBT). "The main idea is to graphically represent the temporally changing evolution of a spatio-temporal object (such as a car or a storm) in a two-dimensional way by its trace. The topological behaviour of such a trace with respect to another object is interpreted and translated into a sequence of predicates, called *development* that can then be used, for example, to query spatio-temporal databases" (Erwig and Schneider, 2003, p. 182). The graphic in Figure 2.8 corresponds to the QBT specification of the *crossFig* predicate of expression 2.9.

Queries in ESTDM

The temporally based queries introduced along with the ESTDM model (Peuquet and Duan, 1995) are location-based queries in a retrospective context. However, they touch mainly changes in a cadastral world and are thus not suited for motion processes.

Indexing moving points

Indexing schemes for storing MPOs are needed to allow fast and efficient responses to motion queries (Agarwal et al., 2003; Pasquale et al., 2003). "The purpose of spatio-temporal indexing is to efficiently support the retrieval of those objects with spatio-temporal extents that satisfy a specified query predicate. The most commonly considered predicate is the intersection with a specified region" (Pfoser and Jensen, 1999, p. 124).

2.5.2 Querying in real time

The aim of moving object databases (MOD) applications is to manage the positions of constantly moving objects in real time (see section 2.3.1). Application fields are digital battlefields, air-traffic control, and mobile communication systems (Basch et al., 1997; Agarwal et al., 2003). Since at query time no updated version of the MPO is available, MOD queries refer to database states that are extrapolated from the last known update. Such queries are called future queries:

- *Future queries.* Queries that refer to future states of database values. For example: "Retreive all airplanes that will come within 30 miles of the airport in the next 10 minutes".

Future temporal logic (FTL)

Querying future states of database values is tightly connected to the underlying data model and its data structures. For example the MOST data model with its dynamic attributes comes along with its own temporal query language called *Future Temporal Logic (FTL)* (Sistla et al., 1997, 1998; Wolfson et al., 1998b,a; Trajcevski et al., 2004). A query example illustrates FTL using the future temporal operator *Until*: "Retrieve all pairs of objects o and n such that the distance between o and n stays within 5 miles until they both enter polygon P" (see expression 2.14).

$$\begin{aligned}&\text{RETRIEVE } o, n\\&\text{WHERE } DIST(o,n) \leq 5\\&\text{Until}(INSIDE(o,P)) \wedge INSIDE(n,P)\end{aligned} \tag{2.14}$$

Sistla et al. (1997) touch upon another critical issue relating to MOD applications for *location based services* (LBS): The time and mode in which the queries are posted. An *instantaneous query* is evaluated exactly when the query is entered, for example "Display the motels within 5 kilometres of my position". In the context of LBS one has to evaluate queries that 'move' along with the moving object. Consider a moving car issuing a query such as "Display motels with availability and costs within a radius of 5 kilometres", and suppose that the car's service system requests the answer to the query to be continuously updated. Such queries are called *continuous queries* (Sistla et al., 1997). In contrast to a continuous query, the *persistent query* considers the whole history starting from initially entering the query. Applied to the example, the persistent query would increasingly collect all motels within 5 kilometres along the whole route.

Instantaneous, continuous, and persistent queries

2.5.3 Querying uncertain positions

Querying MPOs in a prospective way inherently includes uncertainty. Since there is uncertainty about the recent location of the constantly moving objects, a query to a MOD cannot be answered with absolute certainty. The location of the MPO is known with certainty only at the update time. If the degree of uncertainty is controlled, then the danger of incorrect answers to queries due to old data can be reduced. Many approaches allow thus indicating a measure of certainty along with the query result. Wolfson et al. (1999) use probability as a measure of certainty:

Querying uncertain positions

- *Probabilistic Query.* Report not a single object, but a set of objects, each of which have the probability p of answering the query.

A *Probabilisitc Range Query (PRQ)* reports a set of pairs of the form (o, p) where o is an MPO and p is the probability that the object is in the region R at time t. Actually, the algorithm retrieves those MPOs for which p is greater than some threshold. Cheng et al. (2004) extend this approach to answer nearest-neighbour queries. Their *Probabilistic Nearest-Neighbour Query (PNNQ)* reports not a single object that is closest to the object but a set of objects, each of which have the potential of being the nearest neighbour of the query MPO.

PRQ and PNNQ

Relating to the sampling error instead of the preceding motion as the source of uncertainty (see section 2.3.1), Pfoser and Jensen (1999) use explicit spatial concepts to answer another kind of probabilistic queries. For example: "Retrieve the positions of taxis that were inside area A (specified as a rectangle) between

Sampling error as a source of uncertainty times B and C with a probability of at least 30%". As with the sampling error ellipses for the data model, the MPO's positional probability is given by a geometric shape, where a circle represents here the probability function of the worst-case sampling error. An object answers the query if the intersection area of its positional probability circle and the rectangular query given by B and C is 0.3 or higher.

Querying motion: Lessons learned for analysing the motion of points

- Most approaches querying motion focus on location, answering questions about the *where* of MPOs. Trajectory-based queries including higher level motion descriptors such as speed, acceleration, or motion azimuth are rare.
- Querying MPOs involves positional uncertainty, since the actual location of an MPO between two known fixes is normally unknown.

2.6 Visualising motion

Seeing the unseen

The field of *visualisation in scientific computing* (ViSC) emerged from the seminal report by McCormick et al. (1987). The authors define visualisation as a method of computing, transforming the symbolic into the geometric, enabling researchers to observe their simulations and computations. The strength of visualisation for the process of scientific discovery is claimed to lie in its potential to provide profound and unexpected insights (McCormick et al., 1987). This section explores the various contributions of visualising MPOs.

GVIS

Motion has been identified as a key emerging research area in the field of *geographic visualisation* (GVIS): "The computer facilitates direct depiction of movement and change, multiple views of the same data, user interaction with maps, realism (trough three-dimensional stereo views and other techniques), false realism (through fractal generation of landscapes), and the mixing of maps with other graphics, text and sound. Geographic visualisation using our growing array of computer technology allows visual thinking/map interaction to proceed in real time with cartographic displays presented as quickly as an analyst can think of the need of them" (MacEachren and Monmonier, 1992, p. 197).

Visualisation in cartography

For the static cartographic world geographic visualisation provides extensions in at least three significant directions: interaction, animation, and hypermedia (MacEachren and Monmonier, 1992).

These two paragraphs highlight the key elements crucial not only for geographic visualisation but equally relevant for the visualisation of motion (DiBiase et al., 1992; MacEachren and Monmonier, 1992; MacEachren, 1994; Kraak and MacEachren, 1994; Peterson, 1995):

- Temporal maps and temporal cartograms
- Multiple views
- Animation and dynamics
- Interaction

Interaction is a key feature of *exploratory data analysis* (EDA) and will thus be treated in section 2.7 on analysing motion.

2.6.1 Temporal maps and temporal cartograms

Temporal maps

Motion is displacement over time. Thus, maps for visualising motion must be temporal. Kraak and MacEachren (1994, p. 395) define a *temporal map* as a "representation or abstraction of changes in geographical reality: a tool (that is visual, digital, or tactile) for representing geographical information whose locational and/or attribute components change over time."

Mapping trajectories

The simplest way to visualise the motion of an MPO is to map its complete trajectory on a conventional planar map (Turchin, 1998; Broener and Penumarthy, 2003; McGrady et al., 2003; Frihida et al., 2004b; Ramos-Fernández et al., 2004). Kraak and MacEachren (1994) call such maps *single static maps*, – mapping, for instance, the path of a hurricane. Labeling intermediate positions may add temporal information to the track in order to visualise the object's past locations. The symbology and the colour of the trajectory can code motion speed, acceleration, or motion azimuth (Dykes and Mountain, 2003). However sophisticated the symbology may be, with an increasing number of plotted trajectories and larger numbers of MPOs such trajectory maps become confusing.

Trajectories in 3-D

Adding time as a third dimension allows the visual representation of trajectories in 3-D. Thus, the increasing computational power of the last decades gave rise to a diverse set of applications adopting the space-time aquarium data model. Most prominent is the work by Forer's group on visualising (and analysing) student lifestyles and tourism flows in New Zealand (Forer, 1998; Huisman and Forer, 1998; Forer et al., 2004). Kwan (2000) proposes a set of 3-D techniques to analyse disaggregate activity-travel behaviour from travel diary data. As a special feature to facilitate the interpretation of complex interplay of many MPOs she proposes *standardised space-time paths.* To standardise paths they are shifted and rotated so that the home location of all MPOs is the origin $(0, 0)$ and the home-work axis becomes the positive

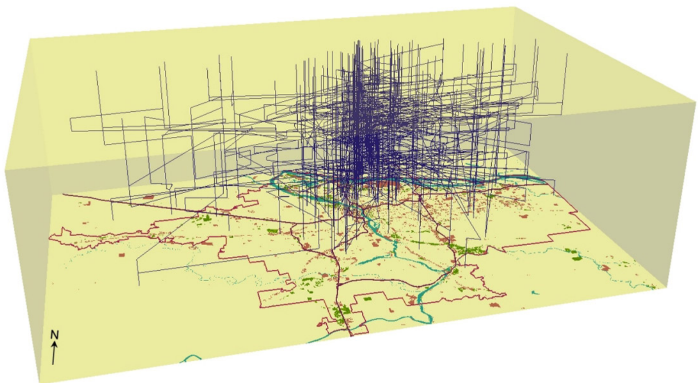


Figure 2.9: The lifelines illustrate individual activity-travel data of three population groups around downtown Portland. Source: Kwan (2000, p. 197).

x-axis. This 'standardisation' allows visual comparison of space-time paths. Finally Kraak and Koussoulakou (2004) present a visualisation environment featuring alternative views, animation, and query functions.

Visual analysis of motion

Alternatively to mapping individual trajectories work has been carried out on static map views summarising motion processes, such as migration processes or the temporal relocation of point entities. Such maps 'summarising' motion may spread over two or three dimensions. Section 2.7 on motion analysis will give a detailed overview of such alternative perspectives of motion.

2.6.2 Multiple views

Changing the perspective

Another key feature in GVIS is the use of multiple views. Multiple views offer the user different insights (and their combination) of the same process. A straightforward multi view visualisation of individual lifelines is appending the 2-D horizontal and vertical projections to the space-time aquarium, representing the grounded trajectory map (xy) and the xt and yt plots respectively (e.g. Kraak and Koussoulakou, 2004). Dykes and Mountain (2003) link a *time view* to the 2-D trajectory map that bar plots the frequency at which the MPOs were logged versus time. This time view allows the identification of *episodes* of spatio-temporal behaviour describing different kinds of acitvity.

2.6.3 Animation and dynamics

Even though animation is not needed to visualise moving objects, the "dynamic display adds an additional set of possible sign-vehicles associated with controls over when and for how long something is displayed. Change in the display over time provides a more direct signification of change in the phenomenon represented" (Kraak and MacEachren, 1994, p. 393). The authors refer to the term *animated maps* to denote to mapping of an event/episode by rapid chronological sequence of static maps or a map that changes dynamically.

Browsing through time

Most static visualisations of motion can be animated by browsing through the temporal dimension and thereby visualising change. Andrienko et al. (2000) propose the *dynamic interval view* with a case study of migrating storks. The interval view shows trajectory fragments during the current interval. The length of this 'worm' shows the speed of movement, with reduction in the length signalling slowing down. In their prototype application for transport demand modelling Frihida et al. (2004b) provide an animated 2-D map view to dynamically visualise individual space-time paths. As a further example Kraak and Koussoulakou (2004) allow the animated 3-D display of MPOs in a space-time aquarium.

Commercial software

Tools for the animated visualisation of motion have recently found their way into commercial GIS solutions. For example, ESRI® offers the ArcGIS™Tracking Analyst extension to visualise tracking data. It features various symbology options and a sophisticated playback manager. However, its power lies almost exclusively in the functionality to define events and to visualise where and when they occur.

MPOs in virtual worlds

Finally computer graphics opens a huge field of applications visualising MPOs. The integration of temporal GIS with computer vision technologies allows highly sophisticated, computer game-like, dynamic visualisations of MPOs in 3-D virtual worlds. Gold et al. (2004), for instance, present a "Marine GIS", where vessels and viewpoints move, and a realistic simulation of the real world view has been designed for navigation and training purposes in marine traffic.

Visualising motion: Lessons learned for analysing the motion of points

- Motion as a geospatial phenomenon can very easily be visualised, be it using static planar or multidimensional, or even animated views.
- However convincing for small sets of MPOs, visualisation of complex motion processes of many MPOs easily overwhelms the perceptional capabilities of the human observer.

2.7 Analysing motion

"Everything is related to everything else, but near things are more related than distant things."

Waldo Tobler, 1970, p. 234

TFL

"*Tobler's First Law of geography* (TFL) is central to conceptions of geographic space at the core of spatial analysis and GIScience" (Miller, 2004, p. 3). According to TFL associated entities may be involved in a causal relationship. "TFL is at the core of *spatial autocorrelation*, that is, quantitative techniques for analysing correlation relative to distance or connectivity relationships. Although correlation is not causality, it provides evidence of causality that can (and should) be assessed in light of theory and/or other evidence" (op. cit.). If TFL is at the core of spatial analysis, so is TFL encompassing time at the core of spatio-temporal analysis: It may not just matter *where* the things are, but also *when* they are there! "Nearness as a concept can be extended to include both space and time" (Miller, 2004, p. 7). Thus, geographic analysis of motion encompasses the following issues:

- aggregation of MPOs (nearness).
- relationships/topology among MPOs (distance, connectivity, direction) (Miller and Wentz, 2003, p. 587).
- path properties (descriptive statistics, in-path autocorrelation)

Featureless space versus environment

These properties of motion can be investigated by adopting two very different perspectives of geographic analysis:

- *Motion analysis in featureless space.* The focus of these approaches lies on the analysis of the motion itself, of its geometric and topologic properties, without considering any influence of the underlying space. Key issues are path properties, changing relationships and topological associations among objects.
- *Motion analysis with respect to environment.* These approaches try to find environmental cues that affect motion. For example, ecologist are highly interested in the interrelation between environmental heterogeneity and individual movement.

This section presents various approaches to analysing the motion of point objects using the headings *descriptive statistics*, *exploratory data analysis* (EDA), and *knowledge discovery in databases* (KDD) and *data mining.*

2.7.1 Descriptive statistics

Step length and turning angle distribution

Descriptive statistics can be applied to all the measures quantifying motion in section 2.2. Given a constant time step, step length represents speed. Step length and turning angle are investigated analysing the motion of animals (Hill and Häder, 1997; Ramos-Fernández et al., 2004). Individual paths or aggregations of many paths can statistically be described deriving distributions of *step length* and *turning angle*, adopting a discrete motion model with straight steps between consecutive fixes. The appropriate statistical description of motion is an important precondition for constrained random walk models covered in section 2.8 on 'simulating motion'.

Analysing the annual cycle

A common variation of the simple step length distribution is the distribution of distances that a group of MPOs moved over one annual cycle. In this case the x axis represents time and the y axis corresponds to distance (Figure 2.10a). Such plots easily illustrate different motion characteristics of seasonally migrating species (unimodal or bimodal distribution) and sedentary species (uniform distribution) (Bergman et al., 2000).

Directional persistence

Autocorrelation in the direction of movement is the key issue investigating turning angle distributions. Trajectories are normally characterised using frequency distributions of discrete classes between -180° and 180° (e.g. Schmitt and Seuront, 2001; Ramos-Fernández et al., 2004). When describing the motion direction a motion azimuth (absolute direction with respect to North) distribution is sometimes preferred over the turning angle. *Radar plots* visualise the turning angle distributions around the compass card in a very illustrative way (Figure 2.10b). Because motion azimuths and turning angles are circular quantities (the angle 180° is the same as the angle -180°, and so are 0° and 360° in the case of motion azimuth), calculating autocorrelation for them requires some care (Turchin, 1998). See Mardia and Jupp (2000) for a comprehensive introduction to the statistics of directional data.

Dispersal capability

For many ecological questions, for example metapopulation dynamics, knowledge about the *dispersal capability* of animals is necessary and thus acquired with extensive empirical and theoretical research (Berger et al., 1999). Berger et al. name three frequently used *linear mobility measures* to describe an individual's motion in ecological field studies:

- *Mean daily movement:* Mean of distances between consecutive fixes divided by the time between those fixes (per day).
- *Maximal distance:* Maximal distance between two fixes (frequently used with mark-recapture studies).

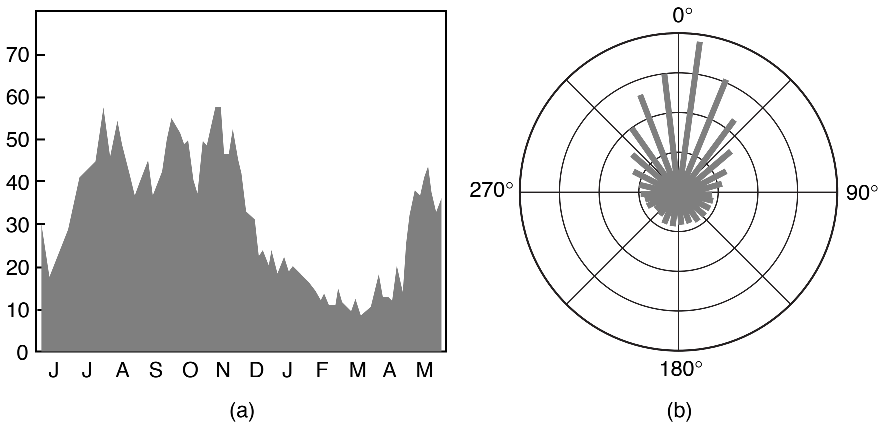


Figure 2.10: (a) Distribution of the mean distance travelled in four days of the individuals of a caribou herd [km]. These migratory animals show a low travel rate in winter, strongly increasing in May starting the spring migration to the calving grounds. (b) Frequency distribution of the successive angles of motion of the same caribou herd, as a radar plot, 0° indicates straight on. Source: Bergman et al. (2000).

- *Mean activity radius:* Average distance between first fix (e.g. capture point) and all consecutive fixes.

Berger et al. (1999) showed with empirical individual-based experiments that the observational interval is the key factor (amongst many others) influencing the accuracy of the three measures.

2.7.2 Exploratory data analysis of motion data

Tracking motion processes generates very rapidly very large datasets. The holistic perspective of descriptive statistics may miss periodic or event patterns as well as locally clustered patterns hiding in the motion data. However, human cognition is very well able to grasp such patterns given the appropriate graphic data representation. The ultimate goal is "to combine the speed, perfect recall, and tireless computational power of modern computing technology with the capability of humans to recognise complex visual patterns, as well as intuition and trained judgement" (Peuquet, 2002, p. 7). Thus, motion analysis requires tools that allow a flexible partitioning of the problem space as well as intuitive and interactive visualisation of the selected subsets to enhance exploration and analysis (Lee and Kemp, 1998). Hence, exploratory data analysis is a further tool which has been used in analysing motion.

Exploratory data analysis (EDA) consists of a collection of descriptive and graphical tools intended to discover patterns in data and suggest hypotheses by imposing as little structure as possible (Tukey, 1977). "EDA is supposed to lead to potentially explicable patterns and is qualitatively distinct from simple descriptive statistics. [...] Modern EDA methods emphasise the interaction between human cognition and computation in the form of dynamic statistical graphics that allow the user to directly manipulate various 'views' of the data. Examples of such views are devices such as histograms, box plots, q-q plots, dot plots, and scatter plots" (Anselin, 1998, p. 78).

EDA

EDA approaches to investigate spatial data must take into account spatial heterogeneity and spatial autocorrelation. *Exploratory spatial data analysis* (ESDA) is then a "collection of techniques to describe and visualise spatial distributions, identify atypical locations or spatial outliers, discover patterns of spatial association, clusters or hot spots, and suggest spatial regimes or other forms of spatial heterogeneity" (Anselin, 1998, p. 79). According to Anselin the conceptualisation of spatial autocorrelation, the fact that locational similarity is linked to value similarity, is central to ESDA. What is defined here for ESDA holds also for a spatio-temporal process such as motion. Hence, incorporating the specific properties of time the concept of ESDA could be adapted for a definition of ESTDA.

ESDA

Interactivity is the key feature of general EDA, and so it is for EDA of motion data (i.e. ESTDA). Interactive maps are a simple yet effective tool for visual data exploration (Andrienko and Andrienko, 1999). This idea is exemplified by the exploration of MPOs trajectories in the 'map view' of the *location trend extractor* (LTE) application presented by Mountain and Raper (2001a) and Dykes and Mountain (2003). The LTE includes the the concepts of *space-time* and *time-space* portrayed earlier (Figure 2.3, page 21).

Interactive mapping

Kwan (2000) proposes a variation of LTE focussing on the attributes of activity-travel data and aspiring to the third dimension. The approach uses a 3-D GIS as a dynamic and interactive data analysis environment for the interactive geovisualisation of activity-travel patterns in transportation research. The user can directly manipulate the attributes of a scene and its features, can change views of the space-time aquarium, alter parameters, query data and see the results of any of these actions easily.

Interactive geovisualisation of activity-travel patterns

Another powerful EDA approach is the transformation of the data in alternative analysis spaces such as e.g. *scatter plots* or *parallel coordinate plots* (see page 51), in order to reduce the di-

Alternative analysis spaces

mensionality and thus the complexity of the motion data to facilitate investigation and analysis. Mountain and Raper (2001a) and Dykes and Mountain (2003) present an approach using density surfaces to overcome the problem of representing large MPO datasets. Their *spotlight view* uses a surface to provide estimates of observation point density at each location. The graphical advantage is that the point density can be represented with a single visual variable, such as colour lightness, and superimposed on further spatial information. Morphometric[2] surface networks derived from the density surface offer furthermore insights about the plotted motion process (Rana and Dykes, 2003). Peaks and ridges on the surface may relate to preferred or regularly used locations and routes, and channels to locations and routes that are avoided. Sadahiro (2002) presents a similar density surface approach for the exploration of spatio-temporal point distributions. In this case the points represent for instance disease occurrences.

TT-plots and RDF

Imfeld (2000) presents a set of EDA methods to analyse MPOs in their environment termed *time plots* and *radial distance functions* (RDF). Time plots plot a geometric property (e.g. distance, direction) against time. Simple time plots have one time axis (T-plots), more complex ones have two time axes (TT-plots). In the TT-δ plot the distances d_{ij} between two consecutive fixes are recorded in the row and column positions of a matrix, corresponding to time steps $t_1, t_2, ...t_n$. Eventually, this gives a symmetric matrix where short distance values indicate that the MPO has not moved much between two consecutive time steps (i.e. it rested), and large values indicate far distances between corresponding time steps. Specific movements produce a specific response in the TT-δ plot, which can then be analysed (Figure 2.11). Up to now, time plots are limited to one or two objects, respectively, and the final interpretation of the movement pattern is performed by human cognition. For RDF plots a series of concentric rings about a location is constructed and the quantity of a property (e.g., forested areas) falling within each ring is measured and plotted against the radial distance. In temporal RDF-plots, RDFs are computed and displayed for consecutive locations moving over time in a plot of distance against time. Like time plots (T)RDFs are limited to a few point objects and exploration is based strictly on human perception.

EDA for group motion

Brillinger et al. (2004) present a set of various EDA techniques applied to a huge collection of VHF telemetry tracked elk and deer.

[2]Morphometric: concerning the characteristics of shape, dimension, or proportion

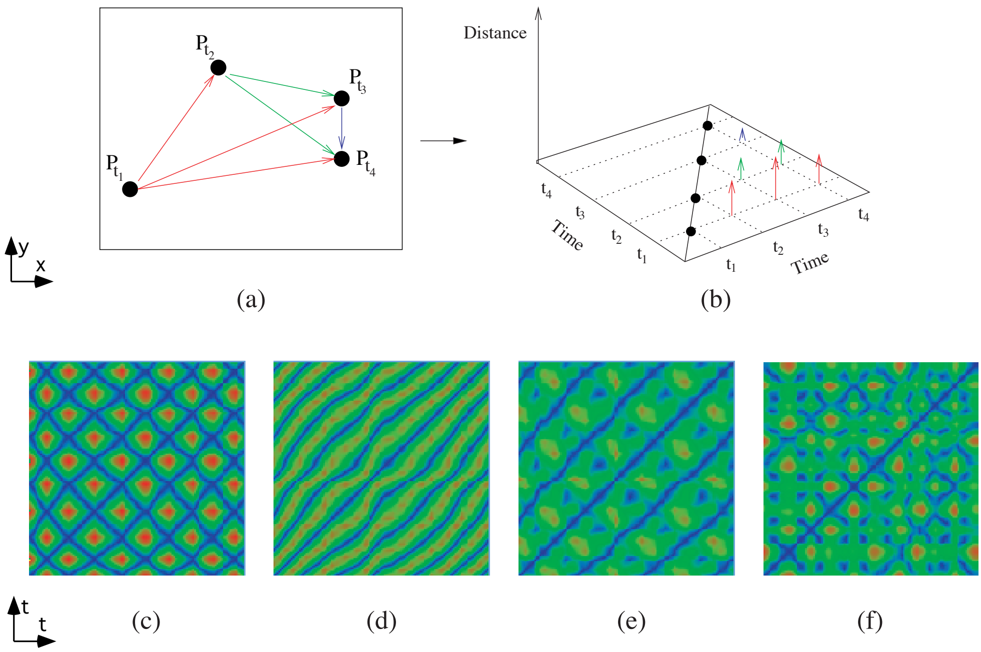


Figure 2.11: TT-δ plots. The construction of the TT-δ plot is illustrated in (a) and (b). (a) shows an example path of an MPO in a planar view with four observed locations. (b) shows the 3-D view of the TT-δ plot construction. The four TT-δ plots below are derived from artificial motion data of an MPO moving (a) forth and back on a straight line, (b) in circles, (c) on an 8-shaped path, and (d) performing a star movement pattern. Source: Imfeld (2000).

Parallel boxplots of the square roots of MPO speeds by hour of the day are used to analyse circadian rhythms. Collapsing all available data for one time of day creates 'temporal transects' suited for descriptive statistics (Figure 2.12). Decomposing the MPO's velocity to cardinal directions using a separate *X-component velocity plot* and a *Y-component velocity plot* provides insights on the directional bias in the joint motion of a group. Finally, *vector fields* address the problem of smoothing and provide a sophisticated overview of the motion of a group of MPOs moving in a distinct area over a distinct time period. Arranged in a regular grid, the length of the vectors is proportional to the estimated speed at that location, its angle corresponds to the estimated motion azimuth. Providing a layer per temporal transect allows the analysis of spatio-temporal variability in a group's motion (Brillinger

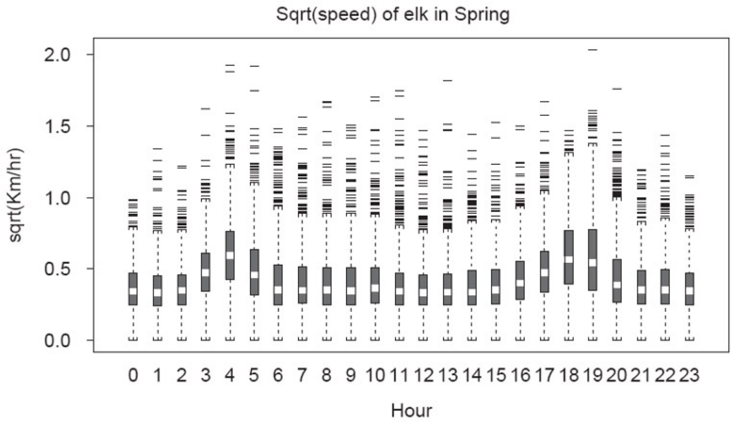


Figure 2.12: This temporal transsects allows the exploratory analysis of a large group of elk. The elk appear much more mobile around 0500h and 1800h. Source: Brillinger et al. (2004).

et al., 2001b,a, 2004).

2.7.3 KDD and data mining

Whereas the early days of Hägerstrand's Time Geography were limited to a data-poor and computation-poor environment, nowadays spatio-temporal analysis is fostered by data-rich and computation-rich environments. Knowledge discovery in databases and data mining are reasonable responses to the huge data volumes in operational and scientific databases covering the motion of MPOs.

Roots in pattern recognition

Today's attempts to derive knowledge from large databases goes back to almost 50 years of research in designing a general purpose machine pattern recogniser in the field of machine learning. "The primary goal of *pattern recognition* is supervised or unsupervised classification" (Jain et al., 2000, p. 4). Watanabe (1985) defines a pattern "as the opposite of chaos. It is an entity, vaguely defined, that could be given a name." In the case of analysing motion one could search for discrete motion events, such as 'sudden parallel move to the north' or 'many MPOs converging to some spot in space'. Based on the above definition of a pattern, Jain et al. (2000, p. 4) gives the following definition of pattern recognition: "(1) supervised classification (e.g. discriminant analysis) in which the input pattern is identified as a member of a predefined class, (2) unsupervised classification (e.g. clustering) in which the pattern is assigned to a hitherto unknown class."

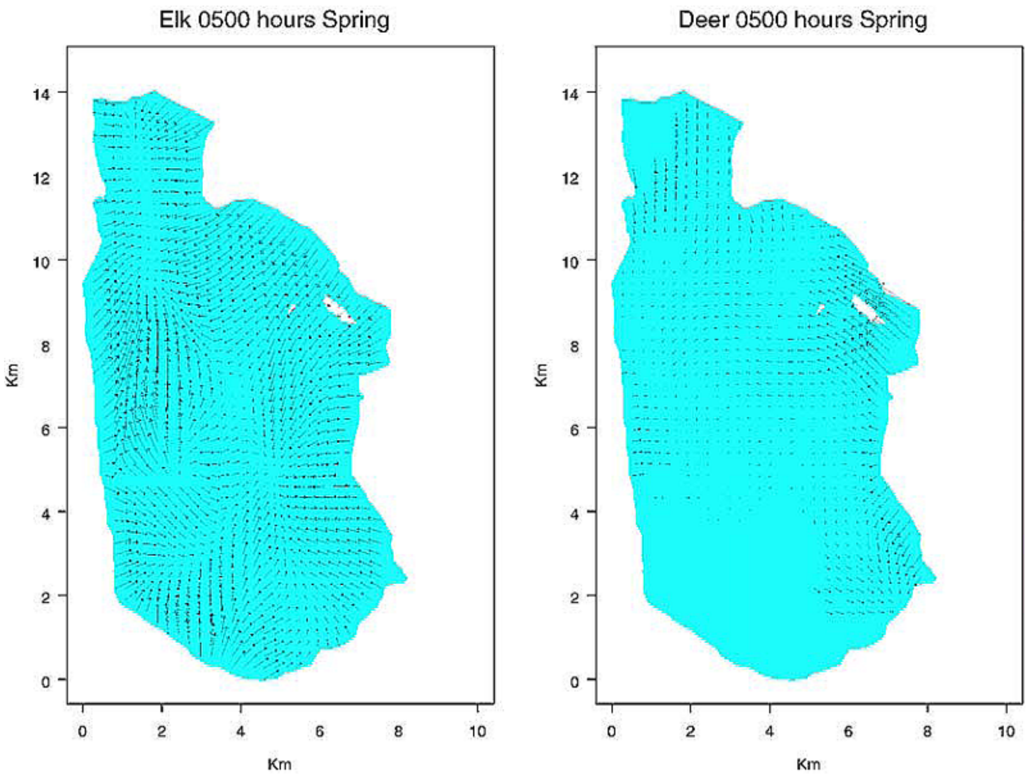


Figure 2.13: The vector fields approximates for every location in space the speed (vector length) and motion azimuth (vector angle) averaged for a large number of MPOs. Source: Brillinger et al. (2004).

Interdisciplinary provenience

"*Knowledge discovery in databases* (KDD) has evolved from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI [artificial intelligence], knowledge acquisition for experts systems, data visualisation, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in the context of large datasets" (Fayyad et al., 1996, p. 39). KDD goes beyond the traditional domains of statistics to accommodate data normally not amendable to statistical analysis. "Statistics usually involves a small and clean (noiseless) numeric database scientifically sampled from a large population with a specific question in mind" (Miller and Han, 2001, p. 5). In contrast, KDD is designed for data collected and stored in many scientific or enterprise databases that are potentially noisy, nun-numeric, and incomplete (Miller and

Han, 2001).

Definition of KDD

Fayyad et al. (1996, p. 40) give a concise definition of KDD: "KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Here, *data* are a set of facts (for example, cases in a database). A *pattern* is an expression in some language describing a subset of the data or a model applicable to the subset. Hence, in our usage, extracting a pattern also designates fitting a model to data, finding structure from data, or, in general, making any high-level description of a set of data. The term *process* implies that KDD comprises many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations. By *nontrivial*, we mean that some search for inference is involved, that is, it is not straightforward computation of predefined quantities like computing the average value of a set of numbers."

Steps of the KDD process

The KDD process is interactive and iterative, involving many decisions made by the user (Fayyad et al., 1996):

1. Identification of the goal of the KDD process
2. Data selection
3. Data cleaning and pre-processing
4. Data reduction and projection (dimensionality reductions, transformations)
5. Selection of data mining method (summarisation, classification, regression, clustering, and so on)
6. Exploratory analysis and model and hypothesis selection
7. Data mining
8. Interpreting mined patterns, possibly returning to previous steps
9. Acting on discovered knowledge, documentation, incorporation of new knowledge in existing theory

Data mining

Data mining is thus just one central component of the overall KDD process. *Data mining* is the application of specific algorithms for extracting patterns from data (Fayyad et al., 1996). The central belief of KDD is that information is hidden in very large databases in the form of interesting *patterns* (Miller and Han, 2001).

Data mining tasks

The various existing data mining algorithms can be categorised as follows: *Segmentation* (clustering, classification trees, artificial neuronal networks), *dependency analysis* (Bayesian networks, association rules), *deviation and outlier detection* (clustering, outlier detection), *trend detection* (regression, sequential pattern extraction), *generalisation and characterisation* (summary rules, attribute-oriented induction). See Miller and Han (2001) for an introductory overview of these data mining tasks.

Pattern interestingness

The KDD community indicates a pattern's interestingness (Padmanabhan, 2004; Silberschatz and Tuzhilin, 1996). Interestingness patterns are divided into those which are objective and subjective. *Objective measures* depend solely on the structure of the pattern and the underlying data. *Subjective measures* depend also on the class of users exploring the data, bearing in mind that patterns that are of interest for one user class, may be of no interest to another class. From a subjective point of view patterns can be interesting because they are unexpected or because the user can act on them. Thus, the *unexpectedness* of a pattern indicates how surprising it is to a user. The *actionability* indicates whether the user can act on it to his advantage.

Geographic knowledge discovery

Miller and Han (2001) identified the unique needs and challenges for integrating KDD into GIScience. They argue that geographic data has unique properties that require special KDD and also data mining approaches. As specific spatial properties they name the geographic measurement framework (Euclidian geometry and topology), spatial dependency and heterogeneity, the complexity of spatio-temporal objects and relationships, and the diversity of involved data types. Hence, they propose the need for specific *geographic knowledge discovery* (GKD) and geographic data mining.

Geographic data mining

"*Geographic data mining* involves the application of computational tools to reveal interesting patterns in objects and events distributed in geographic space and across time. These patterns may involve the spatial properties of individual objects and events (e.g. shape, extent) and spatio-temporal relationships among objects and events in addition to non-spatial attributes of interest in traditional data mining" (Miller and Han, 2001, p. 16).

Bridging KDD and GVIS

MacEachren et al. (1999) bring GVIS and KDD together. Their approach is not explicitly developed for motion data, but it presents an illustrative and typical collection of KDD ideas for multi-dimensional spatio-temporal data, presented here for climate data. The typical KDD ideas could easily be applied for motion data, and the approach is thus discussed here. Three dynamically linked representation forms build their core of the prototype implementation: geoviews, 3-D scatterplots, and parallel coordinate plots (Figure 2.14).

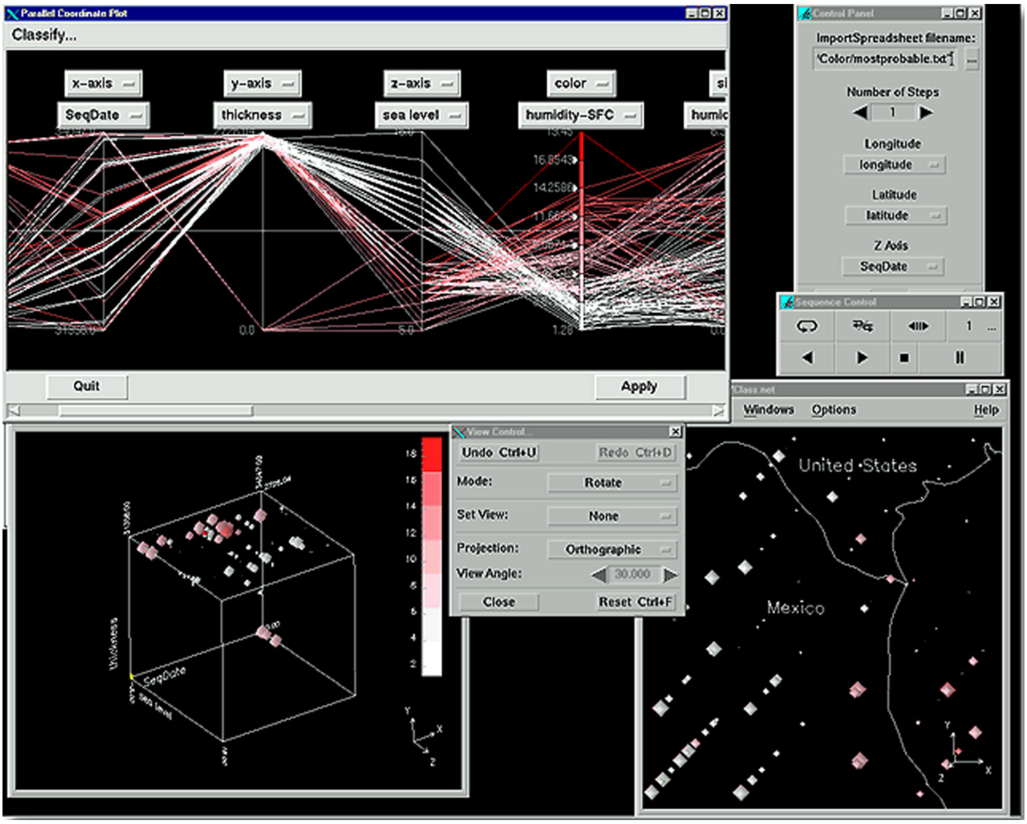


Figure 2.14: Interpretation/evaluation display that includes all three representations for spatio-temporal precipitation data: Parallel coordinate plot (top left), 3-D scatterplot (bottom left), and geoview (bottom right). Source: http://www.geovista.psu.edu/publications/ijgis99/.

- *Geoviews.* 2.5-D windows mapping geographic space (xy) and a third dimension (z) representing elevation or time.
- *3-D scatterplots.* Representations of the relationship between three variables, each plotted as an axis of a cube. This representation uses the central spatialisation metaphor of near = similar, far = different, respectively.
- *Parallel coordinate plots.* Representations that contain several parallel axes, one for each variable in the dataset. The data are distributed along each axis and a line connects individual records from one axis to the next, hence producing a 'signature' for each data record.

Table 2.1: Allen's temporal predicates. Source: Allen (1984).

Relation	Symbol	Pictoral example
X before Y	$<$	XXX YYY
X equal Y	$=$	XXX YYY
X meets Y	m	XXXYYY
X overlaps Y	o	XXX YYY
X during Y	d	XXX YYYYYY
X starts Y	s	XXX YYYYY
X ends Y	e	XXX YYYYY

Interactive manipulations

The three representations are dynamically linked, so that changes executed in one representation are reflected in all others as well. A set of interaction tasks allows the manipulation of the three linked representations: assignment (link data variables to graphic variables), brushing (highlighting single entities or sets), focussing, colour-map manipulation, viewpoint perspective manipulation, and sequencing (use of dynamic variable of order to display one time slice after another).

Building individual space-time path from origin-destination data

Frihida et al. (2004b) propose a KDD approach in the field of transport demand modelling. Their KDD process is designed to extract useful information from an origin-destination survey, i.e. to build individual space-time paths in the space-time aquarium. In this study space-time paths are defined as the ordered sequence of all trips and activities performed by an individual during one day. The ordering uses the concept of temporal topology based on Allen's temporal predicates (Allen, 1984) (Figure 2.1). The KDD process first re-structures the raw origin-destination data into a collection of sequenced individual space-time paths consisting of trips and activities. Second, query and visualisation functions allow one to identify meaningful relationships and processes.

Mining individual trajectories

In a very similar context Smyth (2001) presents a KDD approach to mine mobile trajectories. The overall goal of this research is to gain knowledge from mobile trajectories in order to design better, more scalable, and less expensive location based services. "The central theme of this approach is the mining of stored mobile trajectories, that is, mobile behaviour, in order to predict future behaviour. [...] The fundamental hypothesis is that it is possible to aid travellers in their mobile activities by deriving new information from the traces of their past activities". The author calls this method *stored behaviour-predicted behaviour SB-PB*. The data mining algorithms describe chunks of trajectories using many measurable parameters (such as speed, heading, acceleration), then identify the behaviour of each chunk, and finally store these characteristics in a *behaviour warehouse*, that is assign found motion patterns to archetypical behaviours ready to allocate to new data. For example, a car driver using an in-car mobile device system may benefit from guidance to gas stations, automatically allocated to the stored behaviour 'driving on the highway'.

Analysing motion: Lessons learned for analysing the motion of points

- Many exploratory approaches labelled 'analysis' basically stop at representing the motion and delegate the analytical process to the interpreting analyst.
- Analytical approaches emerging from a cartographic or GIS background adopt a static view comparing snapshots. What is missing are analysis approaches adopting a real process view, where events and processes are analysed and not their instantaneous stamping in static space.
- Many approaches solely investigate the locations of MPOs during a motion process, ignoring the motion properties speed, acceleration or motion azimuth.
- The EDA approaches analysing the motion of elk and deer data presented by Brillinger et al. (2004) are most similar to the objectives of this thesis. The authors do not only ask *when* the animals are *where*, but also which direction they prefer, how fast they move, and whether they show circadian rhythms in speed or motion azimuth.

2.8 Simulating motion

Ants in the garden

The creative act of simulating ants scrabbling around in an artificial garden is at the same time illuminative and fun, – and stunningly easier to code than one could expect (Wood, 2002). The prevalence of the object-oriented paradigm in computer science favoured the diffusion of individual-based motion modelling in many science and engineering fields under many different names and prosecuting many various aims (Batty and Jiang, 1999). However, the principles remain the same. A first requirement is a conceptualisation of the MPOs, – let's say the 'the ants' for the purpose of this introduction. These MPOs may simply be point objects performing some kind of *random walk*. Alternatively the actions of MPOs may be influenced by its neighbourhood as with *cellular automatons*. Or, in an even more sophisticated example, the MPOs may be represented as autonomous software *agents* featuring some behavioural rules, sensing their environment and reacting on it. A second requirement is a conceptualisation of the space accommodating the MPOs and influencing and/or limiting their activities, – 'the garden'.

The idea of *random walk* for a model of completely random motion is a crucial intellectual prerequisite for any reflection on motion. Section 2.8.1 gives a basic overview. The simulation of MPOs is a key feature in behavioural ecology in order to understand population (re-)distribution. Section 2.8.2 gives an overview of the field of *individual-based motion modelling* (IBMM) in ecology. Within GIScience space-time dynamics is receiving increased attention. *Multi-agent systems* (MAS) are an ideal experimental framework to simulate individual actions and decisions in space-time, for instance in the context of urban development or crowd simulation (Section 2.8.3).

2.8.1 Random walk

Random walk

"Within the *random walk* framework, we imagine that an organism travels through the environment by a series of behavioural events, or *moves*. At the beginning of each move, the organism 'makes a decision' as to the duration, speed (or distance), and the direction of its next move. This decision can be influenced by the organism's past experience (in particular, the direction of its previous movement), the local conditions, the conditions at the point of destination (for example, direction and proximity toward a resource patch), and absolute direction (for example, wind direction). Typically (but not necessarily), there is a stochastic element involved in each decision." (Turchin, 1998, p. 128).

Different random walk models

Depending on the interrelation between successive moves, different random walk models can be distinguished (according to Turchin (1998) unless otherwise indicated).

- *Uncorrelated random walk.* The direction of the current move does not affect the direction of the next move, i.e. there is no directional persistence.
- *Correlated random walk.* The direction of the current move does affect the direction of the next move. Step length and turning angle of the subsequent move are drawn from some stochastic frequency distribution, for instance in the case of direction with turning angles concentrated around zero.

- *Biased random walk.* A random walk in which the direction of each move is influenced by an absolute direction, i.e. a long distance attraction (e.g. chemotaxis[3]).
- *Biased correlated random walk.* A random walk in which the direction of each move is influenced by both the direction of the previous move and the absolute direction (i.e. a long distance attraction).
- *Constrained Random Walk (CRW).* Constraints originating from the analysis of real observation data are used to estimate turning angle distributions and step sizes (Wentz et al., 2003).

The application context and the overall goal of the modelling decide which model to choose (see section 2.8.2 and 2.8.3).

2.8.2 Applications in behavioural ecology

Investigate population dynamics

Ecologists are interested in the consequences of movement for population dynamics. The overall goal in investigating motion is to translate individual motion data into a measure of dispersal, or more precisely, population redistribution. The analysis of movement paths is often performed using random walk models to reproduce individual movement. For instance, using the hypothesis of directional persistence, correlated random walk models can be used to investigate relationships between the displacement of MPOs and time (Bergman et al., 2000).

Quantitative analysis of motion

Motion analysis using random walk works in a twofold way. First, a random walk model must be found that best represents the observed motion. Second, various random walk formulae can be used to test qualitative hypotheses about the mode of organism motion. Turchin (1998) provides a well written and understandable introductory textbook to this kind of quantitative analysis of motion.

Random walks in featureless space

Directional persistence

Random walk models allow one for instance to test trajectories for autocorrelation in the direction of movement in featureless space, reflected in the distribution of the turning angles. For a positive autocorrelation in the motion direction, or directional persistence, the turning angles are concentrated around zero (Turchin, 1998, p.

[3]Chemotaxis: The disposition exhibited by certain living cells, or free-swimming organisms, of movement towards or away from certain chemical substances held in solution (Oxford English Dictionary).

135). For example, Bergman et al. (2000) investigated the directional autocorrelation of two differently behaving caribou herds, and could distinguish a migratory and a stationary herd type.

Match specific walk models

Some studies gain insight about the motion characteristics of the species of interest from matching some specific walk model to observation data. Schmitt and Seuront (2001) argue that swimming copepods show *multifractal random walk* behaviour (characterised by a non-linear moment scaling function for the distance versus time), that is trajectories showing alternation between periods of relatively straight swimming and periods of erratic motions including jumps in all three dimensions. Ramos-Fernández et al. (2004) identified so-called *Lévy walk patterns*[4] as an optimal strategy used in searching for scarce resources by spider monkeys.

Mean dispersal rate

Another interesting motion property is the *mean dispersal distance*, a measure to indicate around which distance a population is distributed. Unfortunately animal dispersal is often difficult to observe, consider for instance flying beetles or butterflies, that cannot be observed for more than a few metres. Hence, simulation is used to understand the dispersal process and to generate hypotheses, subsequently tested both by simulation and in the field (Byers, 2001).

Random walks in a heterogeneous environment

"In heterogeneous space one can test whether random walk provides a reasonably accurate description of the mechanisms by which individuals aggregate in response to patchily distributed resources" (Turchin, 1998, p. 157).

Foraging behaviour

One major interest of ecologists using IBMM is the simulation of foraging and ranging in interaction with the underlying vegetation pattern. This may focus on food-related spatial influences (Carter and Finn, 1999; Morales and Ellner, 2002). Some IBMM even include interactions between individuals (Beecham and Farnsworth, 1998), some concentrate on edge effects (Morales, 2002). In many models landscape attributes affect the individual's behaviour. This may be subject to relatively simple rules, such as "if food intake of a cell is below threshold, then move to surrounding cell with highest food value" (Turner et al., 1994). Still other motion models allow estimation of measurement error and process noise that are inherent in animal trajectory data (Jonsen et al., 2003).

Directional bias

Furthermore behavioural ecologists investigate whether paths are afffected by an external bias, e.g. a food source at a special

[4]Trajectories which are composed of self-similar jumps.

location. Whereas CWR assumes an internal directionality (the current move direction is influenced by the previous one), biased random walk models account for an external directionality, i.e. an absolute (compass) direction. Again, the random walk models can be used to match observed paths with alternative mechanisms of how an external bias can influence motion (Turchin, 1998, p. 163). Hill and Häder (1997) derive macroscopic parameters of the trajectories of swimming micro-organisms showing *gravitaxis*[5] and *phototaxis*[6].

2.8.3 Applications in GIScience

Interpolating tracking data

Wentz et al. (2003) propose the use of a constrained random walk model to complete fragmentary trajectories of field observed primates. The existing fragments are used to derive the motion parameters constraining the random walk model, used to construct linking trajectories between the known end points of the existing fragments. The constraints associated with space-time prism concepts applied are (1) a maximum speed of the MPOs and (2) a given use of time per gap, that is, the MPOs do not end up at their destination instantaneously. Finally, the simulated concatenations assist in calculating the home range and in describing the daily ranging habits for different species. The simulated paths well reflected field observations of the movement patterns of the species.

Generation of synthetic data

Pfoser and Theodoridis (2003) propose using motion simulations to overcome the frequent problem of lacking suitable data for designing and testing novel spatio-temporal data types and access methods. They present several examples of how real-world movement characteristics can be translated into appropriate parameters of a spatio-temporal data generator.

Multi-agent systems (MAS)

"As part of the long term quest to develop more disaggregate, temporally dynamic models of spatial behaviour, micro-simulation has evolved to the point where the actions of many individuals can be computed. These *multi-agent systems/simulation* (MAS) models are a consequence of much better micro data, more powerful and user-friendly computer environments often based on parallel processing, and the generally recognised need in spatial science for modelling temporal process" (Batty and Jiang, 1999, p. i).

Agents

"Essentially agents are autonomous entities or objects which act independently of one another, although they may act in concert, depending upon various conditions which are displayed by other

[5]Gravitaxis: moving upwards, responding to a gravity stimulus.
[6]Phototaxis: moving towards a light source.

agents or the system in which they exist" (Batty and Jiang, 1999, p. 3). Franklin and Graesser (1997) define an autonomous agent as "a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future." Regarding the simulation of motion sensing the environment may inform the MPO where it is in absolute space, where it is relative to other MPOs, and where it is with respect to attributes in an environment. A set of rules governs the way in which the agent reacts on these findings leaving its trace in space-time.

Simulate spatial crowd behaviour

Batty et al. (2003) propose the usage of agent-based models to model small-scale spatial events such as carnivals or street parades. The overall goal of this research is to address congestion and safety problems associated with the crowd behaviour at such events.

Modelling object-environment relations

Finally, Westervelt and Hopkins (1999) simulate mobile individuals in dynamic landscapes with cellular automatons. Therefore they implemented an entity based object-oriented simulation environment for MPOs (animals) which are able to collect and store information about their dynamically changing environment in order to adjust against their behavior. The focus of this system is on motion and interaction modelling rather than analysis. However, this is an excellent example of handling object-environment interaction. The approach is limited to small amounts of objects.

Simulating motion: Lessons learned for analysing the motion of points

- Random walk and agent-based models allow the generation of synthetic motion data of any desirable spatial and temporal granularity.
- Due to its high flexibility synthetic motion data is ideal to evaluate and test analytical approaches, performing numerical experiments such as Monte Carlo Simulations or sensitivity analyses.

2.9 Summary

This chapter provided a review of the latest literature covering the motion of point objects. The review put an emphasis on GIScience, deliberately side stepping to computer science, statistics, exploratory analysis and ecology. The references were discussed and thus structured adopting eight separate perspectives, namely capturing, quantifying, modelling, formalising, querying, visualising, analysing, and finally simulating motion.

The chapter could report on a considerable amount of work on modelling moving entities. Major advances could also be highlighted in designing and querying database systems adopted to cover motion. Well researched is furthermore the representation of motion, i.e. visualisation for exploratory analysis approaches. However, this literature review reveals that convincing ideas for analysing motion remain sparse, mirrored in the fact that hardly any motion analysis concept found its way into commercial GIS software. To conclude, the field of analysis techniques for a better understanding of motion processes remains wide open.

Even though the research questions of this thesis are located in the introductory chapter before the state of the art for structural reasons (see page 7), they are of course highly influenced by the lessons learned studying the literature.

Chapter 3

Methods & results

> *"A pattern is the opposite of chaos; it is an entity, vaguely defined, that could be given a name."*
>
> Santosi Watanabe (1985, p. 2)

The main components of this thesis are the four constitutive publications. As a comprehensive access guide for the reader, this chapter sets the frame of this piece of research and summarises rationale, methods, results, and contributions of each of the four publications[1]. This scientific summary does not substitute the reading of the full papers attached to this volume, but rather positions the respective research packages in a broader context and offers the reader a structured account of the research done.

Coauthors The work presented in the following papers is original research undertaken by the author of this thesis, first author of all included publications. Dr. Stephan Imfeld and Prof. Dr. Robert Weibel contributed to the publications in their role as supervisors of this thesis. The contributions of further external coauthors is declared if there is need.

[1]The margin text in this chapter is used to refer to the papers in their originally published form, attached starting at page 133.

3.1 The concept of relative motion

Laube, P. and Imfeld, S. (2002). Analyzing relative motion within groups of trackable moving point objects. In Egenhofer, M. J. and Mark, D. M., editors, *Geographic Information Science*, volume 2478 of *Lecture Notes in Computer Science*, pages 132–144. Springer, Berlin-Heidelberg, DE.

Objective 1: *This thesis shall identify, characterise, and categorise a set of basic motion patterns in the lifelines of groups of MPOs, with motion patterns being a predefined formalised search template of motion attributes such as speed, change of speed, or motion azimuth.*

3.1.1 Rationale

This paper reports on conceptual GIScience developing an integrated GKD approach for analysing geospatial lifelines of groups of moving point objects. The overall goal is to conceptualise a flexible framework to find, quantify, and visualise user-defined motion patterns in the trajectories of groups of MPOs. Therefore the paper discusses the identification, characterisation, and categorisation of generic motion patterns. In the course of this thesis, this first paper built the entry point, describing the conceptual framework underlying this research.

3.1.2 Methods

The key concept of the proposed GKD approach and hence this thesis is to compare the motion parameters of MPOs over space and over time, giving the concept its name: REMO (RElative MOtion). Suitable geospatial lifeline data consist of a set of MPOs each featuring a list of fixes, tuples of (x, y, z, t).

REMO is a classical KDD process (Fayyad et al., 1996), focusing on the basic KDD steps *data reduction and projection* and *data mining*, respectively. Since the application field is geographic, the KDD process is referred to as *geographic knowledge discovery* (GKD). The first step consists of a transformation of the geospatial lifeline data into a 2.5-dimensional analysis matrix featuring a time axis, an object axis and motion attributes (i.e. speed, change of speed, and motion azimuth). It is assumed that specific motion behaviour and interrelations among the MPOs are manifested as patterns in the analysis matrix. Thus, as a second step, predefined formalised relative motion patterns (so-called *REMO patterns*) are matched on the analysis matrix. For the second

Figure 1, p. 134.

step the potential user, having knowledge about the investigated process, parameterises the patterns to be found.

3.1.3 Results

The paper defines relative motion patterns as sets of cells in the analysis matrix, having an extent in time and/or across the objects. The paper proposes and describes a set of generic relative motion patterns. The simple patterns are termed *constancy*[2], *turn, concurrence, opposition*, and *dispersion*. The complex patterns are *trend-setter, independent, propagation, group turn* and, *group concurrence*. Figure 2, p. 138.

In the 'Test Data Sets' section the paper illustrates the GKD process using two case studies. The Porcupine Caribou Herd Satellite Collar Project represents an animal tracking project in wildlife biology. The second case study covers abstract entities moving in an abstract data space, namely Swiss political districts moving in a space representing ideological preferences in Switzerland. In both case studies basic REMO patterns such as constancy, concurrence, and trend-setter behaviour could be identified using visualisation of the REMO matrix and qualitative analysis. Figure 3, p. 140. Figure 4, p. 142.

3.1.4 Contributions

The main contribution of this paper is the proposal of a new GKD approach for the spatio-temporal analysis of geospatial lifelines. The paper therefore identifies and categorises a set of generic relative motion patterns and conceptualises a GKD procedure based on the combination of a (re-)projection of the tracking data with syntactic pattern detection methods. The paper underlines the universal applicability of the approach using two very different case studies for illustration and qualitative analysis. Finally, the paper provides the conceptual framework underlying the remainder of the publications in this research.

[2]Note: Reacting to a reviewer's suggestion, the term *constance* used in the original paper was replaced by *constancy* in subsequent work. However, both terms represent the same motion pattern.

3.2 Formalising and detecting motion patterns

Laube, P., Imfeld, S., and Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668.

Objective 2: *This thesis shall develop a knowledge discovery approach integrating techniques to formally describe motion patterns and algorithms allowing detection of these patterns in case study data.*

3.2.1 Rationale

This publication presents research implementing and extending the basic REMO concept presented in section 3.1. The main objective of this stage is to demonstrate the feasibility of the proposed concept, focusing on the aspects *formalisation* and *implementation*. This paper intends to develop a simple and generic pattern description formalism to describe relative motion patterns and algorithms to detect instances of these patterns on geospatial lifelines. It shall furthermore address a series of implementation related issues such as developing an object-oriented class design for the REMO GKD approach, sampling lifelines at different granularities adopting different interpolation and aggregation strategies, and introducing a user-friendly graphical user interface (GUI) to control and evaluate the GKD process.

3.2.2 Methods

This paper develops the computational framework to test and improve the REMO GKD approach presented in section 3.1. This includes the development of an object-oriented class design to model groups of MPOs and to mine their trajectories for motion patterns. A special emphasis is put on handling incomplete and variably sampled lifelines as well as on the problem of different temporal analysis granularities.

The REMO GKD approach follows the syntactic pattern detection approach since simple patterns serve as primitives for the construction of more complex patterns (Jain et al., 2000). In the data mining process expected patterns are defined in advance based on expert knowledge about the investigated motion process and subsequently instances of the described patterns are matched in the input data. This procedure requires the development of a description formalism for motion patterns on lifelines. Conse-

Figure 6, p. 653.
Table 1, p. 649, table 2, p. 650.

quently, the paper proposes a REMO pattern description formalism adopting elements of the commonly used regular expression formalism (regex) as well as of basic mathematical logic.

3.2.3 Results

This paper tested the REMO GKD approach and its prototype implementation with data from (a) football players tracked on the pitch and (b) data points in an abstract ideological space. With the soccer team typical moves such as straight pushes of attacking offensive players as well as coordinated defensive withdrawal could be identified and quantified. An example for moving abstract data points were the tracks of Swiss political districts displacement in an ideological space with respect to their temporally changing voting behaviour (Hermann and Leuthold, 2001, 2003). As a most demonstrative result trend-setting districts could be identified that anticipate many years in advance political divergence towards the political left and right, respectively. These illustrative results show that the proposed methodology is able to extract a set of valid, novel, useful and understandable motion patterns in geospatial lifeline data.

Figure 7, p. 655.

Figures 11 and 12, p. 658.

3.2.4 Contributions

As a first main contribution a simple yet user-friendly formalism has been developed that allows specification of motion patterns for the syntactic pattern detection approach. Tests with various test data illustrated the approach's generic nature and applicability. As a second main contribution this paper provided the proof of concept for the REMO GKD approach. The conceptual ideas presented in section 3.1 could successfully be implemented in a prototype implementation coded in Java. The implementation work used an object-oriented class design allowing the derivation of motion attributes at variable temporal granularities from imperfect lifelines using the so-called *detached attribute functions*. With respect to computation issues this paper delivered the required data mining algorithms. Overall this paper showed that syntactic pattern detection and spatio-temporal data mining in general are promising analysis techniques to handle the current increase of spatio-temporal data.

Figure 5, p. 652, Figure 8, p. 655, Figures 9 and 10, p. 657.

Figure 3, p. 645.

3.3 Motion patterns in absolute space

Laube, P., Van Kreveld, M. and Imfeld, S. (2004). Finding REMO - detecting relative motion patterns in geospatial lifelines. In Fisher, P. F., editor, *Developments in Spatial Data Handling*, Proceedings of the 11th International Symposium on Spatial Data Handling, pages 201–215. Springer, Berlin-Heidelberg, DE.

Objective 3: *This thesis shall develop means to describe and algorithmically detect forms of temporally coordinated and spatially directed group motion patterns in trajectories of MPOs, such as flocking, converging or diverging.*

3.3.1 Rationale

The REMO GKD approach presented in the sections 3.1 and 3.2 focused purely on properties describing the motion of MPOs, explicitly excluding the absolute positions of the MPOs. Excluding the absolute positions is a valid approach to reduce the complexity of the motion process. However, MPOs do not disclose their complex interrelations solely in their motion properties but also in changes of their arrangement in absolute space. This paper reports on the extension of the REMO GKD approach to incorporate the (dynamic) arrangement of the MPOs in absolute (geographic) space. The proposed spatially constrained patterns are able to describe flocking behaviour as well as convergence/divergence processes in groups of MPOs. The paper discusses the geometric properties of the formalised patterns with respect to their efficient computation.

3.3.2 Methods

The methodological core of this paper is the integration of spatial data handling, computational geometry and KDD. The paper resumes the data mining approach of sections 3.1 and 3.2 but proposes new patterns incorporating spatial constraints for the fixes of the MPOs in absolute space. Proximity measures known from the field of spatial data handling are used to express proximity relations between MPOs. The second part of the paper develops the pattern detection algorithms needed to identify spatially constrained REMO patterns. The efficient detection of the spatially constrained patterns makes use of geometric algorithms. Higher-order Voronoi diagrams are used to test spatial proximity of the fixes of a pattern. The overlay area of half-strips, representing the MPOs projected motion vectors, is used to detect convergence. A

Figure 3, p. 209

detailed efficiency analysis of the proposed algorithms completes the paper.

3.3.3 Results

The primary results are the formalisation of the spatially constrained REMO patterns *track*, *flock*, *leadership*, *convergence / divergence*, and *encounter / breakup* and the development of the associated algorithms to detect these patterns. The algorithm efficiency analysis showed that the introduced patterns can be detected within reasonable time. Many simple patterns can be detected in close to linear time. Some more complex patterns (e.g. encounter) are quite expensive to compute, but still feasible to compute in off-line analysis in the case of datasets consisting of several hundreds of objects.

Figure 2, p. 207

3.3.4 Contributions

This paper successfully transferred the relative motion concept to absolute space. The integration of spatially constrained motion patterns opens up a promising research field to analyse geospatial lifelines. With the aggregation patterns *convergence / divergence*, and *encounter / breakup* respectively, a new kind of motion behaviour pattern could be identified that considers both the relative and absolute positions of MPOs. These patterns are intrinsically dynamic and exist neither solely in space nor solely in time, they only exist in a dynamic view of the world. The aggregation patterns furthermore illustrate the need to separate the concept of static arrangement from dynamic aggregation. A set of MPOs can very well perform the *aggregation* process and converge for a long time without building the static *arrangement* of a cluster. Conversely, MPOs moving around in a cluster may not converge at all. Even though convergence and clustering are often spatially and/or temporally related, there need not be a detectable relation in an individual data frame under investigation. Once again, this shows the limits of the snapshot view of the world.

Coauthors This work emerged from the cheese cake session at Dagstuhl Seminar No. 03401 on "Computational Cartography and Spatial Modelling", 28th September – 3rd October 2003, Schloss Dagstuhl, Germany. Dr. Marc van Kreveld, currently assistant professor at the Institute of Information and Computing Science at Utrecht University, contributed to this work in particular with respect to the integration of computational geometry within the REMO framework.

3.4 Evaluation of the REMO concept

Laube, P. and Purves, R.S (submitted 2005). Evaluating motion pattern techniques in spatio-temporal data. *Computers, Environment and Urban Systems*, Submitted April 2005.

Objective 4: *This thesis shall develop statistical measures and adopt Monte Carlo techniques to quantify the relevance and interestingness of the proposed motion patterns, to distinguish random noise from information in the knowledge discovery process.*

3.4.1 Rationale

It has been recognised in the KDD literature that discovery systems can find a glut of patterns, most of which are of no interest to the user (Silberschatz and Tuzhilin, 1996; Padmanabhan, 2004). Within the REMO GKD framework introduced so far, the potential user searches lifeline data for instances of pre-defined motion patterns, which are constructed on the basis of the user's knowledge about the motion process under study. However, the user has no means by which to estimate the relevance of the extracted patterns. Hence, the final paper addresses the question of evaluating the relevance of the motion patterns, described through their *interestingness*. The paper uses again the Porcupine caribou data and the Swiss districts moving in the abstract data space introduced in section 3.1.

3.4.2 Methods

Silberschatz and Tuzhilin (1996) propose *unexpectedness* as a measure of interestingness of patterns. They argue that patterns are interesting because they contradict our expectations, given by our system of beliefs. One option to capture the beliefs is to formulate a statistical hypothesis. The degree of belief is then defined as a significance level for which the statistic for a certain test is on the borderline of acceptance of the hypothesis. The main drawback of this statistical approach to indicate unexpectedness is that not any belief can be formulated as a testable hypothesis (cit. op.). For this reason the proposed evaluation technique replaces the statistical hypothesis with Monte Carlo simulations, choosing the detour of numerical experiments to learn about "the expected" from the stochastic properties of the simulations.

The strategy for evaluating the REMO GKD approach is twofold. First, Monte Carlo simulations are used to generate a population

Figure 9, p. 13.

of simulated lifelines. The concept of *constrained random walk* (CRW) is used to simulate lifelines which have similar statistical properties to the observed data. The constraints are given as frequency distributions of step length and turning angle derived from the observation data. In a second step the number of patterns found in the synthetic data are compared with the number found in the observational data. The underlying assumption is that those patterns which appear to be outliers from the stochastic properties of the simulations are those which one can attach some initial interestingness to, prior to further investigation by the user.

Figure 8, p. 13.

Figure 4–7, pp. 10–11.

To estimate the influence of the characteristics of the raw data and the configuration of the pattern matching process, a series of numerical experiments was performed systematically varying the pattern matching parameters, that is the extent of the searched pattern as well as the attribute granularity.

Figures 10, p. 14

3.4.3 Results

The MC experiments with both the Caribou and the Swiss district data point out that performing the REMO GKD process a lower pattern length/width threshold must be adopted depending on the classification schema, the finer the attribute classification, the lower this threshold has to be set. Patterns found with extents below this threshold cannot be considered as unexpected.

The experiments performed on the motion azimuth of migrating caribou resulted in two main findings. Constancy patterns found in the observation data are no interesting anomaly but rather expected from the simulation data. Having in mind that the temporal granularity δt of two weeks is rather coarse with respect to the total length of the migration periods, it is not surprising that no significant constancy patterns could be found. The number of concurrence and trend-setter matches found in the observation data, in contrast, is for certain pattern extents significantly above the expected number from the simulations. This result allows one to assign interestingness to the the instances of concurrence and trend-setter found in the observation data.

Figure 11, p. 15.

Figure 12, p. 16,

Figure 13, p.17.

Figures 18 and 19, p. 21.

Very similar thresholds of interestingness could be identified for concurrence in motion azimuth of the political districts. Interestingly, following the results of the numerical experiments, interestingness could only be assigned to some of the concurrence patterns emerging from a political left-right divergence discussed in section 3.2.3. While a concurrence pattern including the concurrent motion of more than 45 districts appeared interesting, the concurrence of only 18 districts did not.

Figure16, p. 19.

Figure 22, p. 24.

3.4.4 Contributions

This paper presents a method to assess the interestingness of motion patterns that are detected through the REMO GKD approach. The numerical experiments underpin first the hypotheses that the configuration of the pattern detection process heavily influences the number of patterns found. Second, in many cases, found patterns must be classified not interesting, since the emerge from the simulation data with the same frequency as observed. Hence, the evaluation method introduced tools to help identify useful configurations of pattern detection parameters.

Coauthors This publication presents joint research undertaken with Dr. Ross Purves, currently lecturer at the Department of Geography at the University of Zurich. He contributed to this work with his rich experience on numerical experiments and the modelling of complex environmental processes.

3.5 Summary

This chapter summarised each of the four publications contributing to this thesis. section 3.1 reported about the development of the basic concept of relative motion and a set of generic motion patterns. Section 3.2 described work undertaken to give a proof of concept for the REMO GKD approach, focussing on formalising and algorithmically detecting these patterns. Section 3.3 summarised work extending the list of motion patterns by a set of spatially constrained patterns. Finally, section 3.4 gave an account of a technique to assess the interestingness of detected motion patterns.

Chapter 4

Discussion

> *"Auf die Idee, dies als Trendsignal zu deuten, kommt Geri in diesem Frühstadium der Entwicklung noch nicht. Erst später, als auch Robi Meili am Stammtisch fehlt, obwohl er weder krank noch in den Ferien ist, überfällt Geri Weibel eine erste Ahnung des Undenkbaren. Könnte es sein, dass das Mucho Gusto...?[1]"*
>
> Aus *Richtig leben mit Geri Weibel*,
> Martin Suter (2001, p. 55).

This chapter provides a synthesis of the four individual publications, bringing together the individual strands of the papers into a holistic discussion. The discussion integrates the work presented in the four publications listing first strengths and weaknesses of the proposed GKD approach, naming the approach's applicability and limitations, and finally recalls the research questions from section 1.2 motivating this thesis. Synthesis

4.1 Evaluation of the REMO GKD approach

This evaluation section identifies strengths and weaknesses of the proposed REMO GKD approach. Aspects considered as *strengths* are advantages, capabilities, competitive advantages, innovative aspects, efficiency, unique experiences, and originality, disadvantages of proposition, gaps in capabilities, or own known vulnerabil-

[1] *"In this early stage of the development, Geri did not recognise it as a trend signal. Only later, when also Robi Meili did not show up at the regulars' table, even though neither being sick nor on holidays, Geri got struck by a first presentiment of the unthinkable. Could it be that the Mucho Gusto was...?"* Geri Weibel is the urban trend scout of the Swiss columnist Martin Suter.

ities are viewed as *weaknesses*. Table 4.1 provides a first overview of the strengths and weaknesses which are subsequently discussed.

Table 4.1: Strength and weaknesses of the REMO GDK approach

strengths		weaknesses	
$\oplus$	True integration of space and time	$\ominus$	Dependency on expert knowledge
$\oplus$	Simplicity and understandability	$\ominus$	Imperative of discretisation
$\oplus$	Universal applicability	$\ominus$	Modifiable temporal unit problem
$\oplus$	Extensibility of approach		

4.1.1 Strengths

True integration of space and time

The REMO GKD approach integrates spatial and temporal analysis. The approach is intrinsically spatio-temporal as is its motivating phenomenon motion. The proposed motion patterns are neither purely spatial nor purely temporal but span space *and* time.

Simplicity and understandability

REMO GKD is simple and ultimately understandable. The REMO patterns are descriptive, based on well-known everyday motion events, and thus easy to learn and apply for potential users.

Universal applicability

The phenomenon of motion is widespread in many different fields. The moving point metaphor is the minimal representation for motion, i.e. its narrowed essence. First, as will be shown in section 4.3.1 the REMO motion patterns based thereon can be considered as generic, potentially found in the motion records originating from a large variety of application fields. Second, the REMO DKD process is also generic. The sequence of defining patterns, their subsequent detection in an analysis matrix and the final quantification of the data mining results is applicable to a wide range of motion phenomena.

Extensibility

The REMO pattern description formalism follows the hierarchical approach of syntactic pattern recognition, providing a set of simple pattern primitives to compose complex patterns. This strategy allows the composition of arbitrary motion patterns satisfying unforeseen needs of potential users.

4.1.2 Weaknesses

Dependency on expert knowledge

The REMO GKD approach requires preliminary expert knowledge about the process under investigation. Potential users, scientific experts in fields such as biology, geography, or sociology, must compose REMO patterns potentially lurking in the geospatial lifelines. Thus, they must have an idea of what they expect to find. In most scientific projects this knowledge is available, but nonetheless, such dependency on expert knowledge makes analysis subjectively dependent on the experts' skills.

Imperative of discretisation

The REMO GKD framework allows one to use arbitrary temporal analysis granularites. Yet, however fine this granularity may be, it still implies a discretisation of the temporal axis.

'Modifiable temporal unit problem'

Another crucial issue arises which can be compared to the modifiable areal unit problem (MAUP) (Openshaw, 1984). Discretising lifelines we consider different temporal aggregations instead of spatial aggregations as with the classical MAUP. The arbitrary aggregation may emerge from the mapping of potentially irregularly sampled fixes on to the regular REMO matrix, determined by an analysis granularity and a starting time. If the temporal units were specified differently, we might observe very different patterns and relationships.

Simplicity dilemma

The strongest advantage of the REMO approach may be at the same time it's most dangerous weakness, – an issue that hence could be put on both lists. The patterns are easy, convincing, near to real life, and easy to construct, – and some times easy to find. However, like every other knowledge discovery approach adopting data mining activities, REMO GKD is able to find patterns that appear significant but, in fact, are not (Silberschatz and Tuzhilin, 1996; Laube and Purves, 2005). "Data mining is a legitimate activity as long as one understands how to do it correctly, data mining carried out poorly (without regard to the statistical aspects of the problem) is to be avoided" (Fayyad et al., 1996, p. 40). Thus, with its user friendly simplicity the approach may encourage careless use. To overcome this dilemma, this thesis proposed a technique to assess the interestingness of the mined patterns (Laube and Purves, 2005) (see section 4.3.7).

4.2 Applications and limitations for the REMO GDK approach

The REMO GKD approach is designed to investigate the motion expressed in groups of many MPOs. However, experiments involving case studies from diverse fields made clear that the REMO

GKD approach is not suited for all possible kinds of motion data, and *vice versa* not all data mining strategies provided within the REMO are suited for specific case studies. This section shall discuss for what kind of research tasks the REMO approach is suited and which limitations should be considered.

4.2.1 Finding suitable motion data

REMO's crux with animal data

Wildlife biology was a great motivator for the development of the REMO GKD approach. It is very easy to conceptualise convincing motion patterns potentially lurking in trajectories of tracked animals. Unfortunately it is at the same time very difficult to actually find motion patterns in real observation data. A first explanation for this dichotomy may be found in the (non-)availablility animal tracking data. The REMO GKD approach has considerable requirements on the nature of the geospatial lifeline data. The strongest constraint is the need for contemporaneous fixes. In addition, even if one finds data covering, for example, a dozen individuals at the same time, these animals usually are each representatives for different and remote groups rather than members of a single group. A second reason may be that, for example, foraging animals simply don't express the kind of patterns proposed with the REMO approach. Who would doubt the quality of a cluster detection algorithm failing to find clusters in randomly scattered data.

In the same boat with the space-time aquarium

As is every other KDD approach REMO GKD is designed to make use of the nearly unlimited computing power of today's computers. The approach is thus especially suited for large numbers of MPOs and very long tracking periods. Unfortunately it is still hard to find case study data showing these properties. Actually, the applications used in the thesis provide rather small numbers of MPOs, ranging from some individuals to over one hundred objects. Tracking large numbers of individuals is expensive and laborious. Furthermore, *geoprivacy* issues may limit data availability. The main concern with respect to spatio-temporal GIS is its potential for a rapid integration of spatial information and personal information (Dobson and Fisher, 2003). Ethical concerns and objections may additionally constrain the availability of geospatial lifeline data in the years ahead. Hence, the approach could share for some time the destiny of the space time aquarium, remaining an elegant and promising concept, yet suffering from a lack of true applications.

Satellites or taxi cabs?

The data model chosen for REMO GKD is optimised for data emerging from GPS systems or VHF telemetry systems, usually providing system inherent regularly sampled fixes. Event-delimited data originating from a MOD applications, such as a taxi cab fleet management system, are less suited for REMO GDK. Such event-based data featuring long static periods and rare updates are not suited for the basic idea of REMO, since it requires periodically (inter-)relating the motion properties of MPOs.

4.2.2 REMO's relation to space

Unconstrained motion

The REMO GKD approach assumes that the MPOs move without constraints in a featureless space. In many cases this is an oversimplification of reality, well knowing that for instance animals actually frequently follow corridors or migrate in valleys rather than on ridges. However, such special constraints are application specific and the expert's intimacy with his/her data and process must avoid futile analysis tasks. Still, the application of the REMO GKD approach is not suited for the analysis of tracking data of MPOs moving on a network. For instance, investigating motion azimuth of cars moving on a highway network potentially reveals something about the structure of the traffic network but little about the behaviour of the MPOs.

Mike Worboy's quest for space

After a presentation of the principles of the REMO GKD approach at the Dagstuhl Seminar "Computational Cartography and Spatial Modelling" (28th September – 3rd October 2003, Schloss Dagstuhl, Germany), Michael Worboys asked the question: "Where is space in the REMO set-up? Instead of motion descriptors any other attribute changing over time could be used for the data mining." This objection is legitimate. The REMO concept introduced in Laube and Imfeld (2002) is strongly linked to the motion properties speed, change of speed, and motion azimuth. Nevertheless, investigating motion as a process, it is a valid assumption that spatio-temporally changing properties describing motion may code process knowledge. However, as a direct reaction to this objection, Laube et al. (2004) successfully extended the spatial sensitivity of the REMO GKD approach.

4.2.3 Granularity issues

Sampling rate vs. analysis granularity

A first critical issue relating to granularity is the interplay of sampling and analysis granularity. Undersampling a lifeline causes information loss, oversampling may drown out the track's signal and may even feign autocorrelation between successive moves (Laube et al., 2005). To avoid undersampling collecting data at the highest sampling rate possible is a simple yet effective strategy. The problem of oversampling is more subtle. One strategy is to resample the tracks at an increasingly coarser granularity, until autocorrelation between the moves disappears (see Turchin (1998, p.

130) for details on how to avoid oversampling when discretising tracks). In order to avoid semantic mismatches the GKD process must be performed at an adequate granularity. Choosing an analysis granularity coarser or equal to the fix sampling rate limits the danger of oversampling. Hence, the sampling rate of the input data, normally given and not changeable by the user, limits the analysis. The phenomenon under investigation further suggests suitable analysis granularities. Searching for seasonal migration patterns one should not choose an analysis granularity of hours, introducing noise caused by daily movement patterns irrelevant to the research question.

Different reading of motion patterns at different granularities

Some patterns have to be interpreted in another way depending on the analysis granularity chosen. Take, for instance, concurrence. In a small group concurrence at an analysis granularity of hours may be evidence for in-group decision making. In a group with individuals at remote locations without any contact, a concurrence may be evidence of an inner stimulus, e.g. the triggering of seasonal migrations. Note that this thesis does not attempt to make contributions to research in wildlife biology or behavioural ecology and hence does not try to explain animal or human behaviour. However, this last example shall illustrate that the given application and the analysis task may well impose limitations on the use of the REMO GKD approach. Depending on the kind of concurrence that is to be investigated it may or may not be required that the MPOs of the case study actually move in a spatially aggregated group.

Variable granularities within trajectories

Lifeline data covering long time periods often include variable sampling granularities. For example in an ecological field study investigating caribou behaviour, the most interesting months may be in winter. Thus, for economic reasons the winter months may feature eight fixes a day whereas the summer months may only feature three fixes a day. With the detached attribute functions introduced in Laube et al. (2005) it is possible to derive reasonable interpolations of the motion azimuth at, say, an analysis granularity of one day. For the motion properties speed or sinuosity, by contrast, the interpolation remains an open problem. Consider speed: Having two different granularities (8 fixes/day and 3 fixes/day), and the MPO possibly wiggling around in the first case, just dividing the track length by the time interval will cause severe artefacts. The same is true for sinuosity, again resulting in much higher values at a finer granularity. A general theory of integrating variable granularities in geographic knowledge discovery in motion data remains an open issue.

Attribute classification

Yet another granularity effect is due to the classification of the motion attributes. The number of matched patterns is highly dependent on the attribute granularity of the pattern matching process (Laube and Purves, 2005). The phenomenon under study gives a suitable analysis granularity (e.g. weeks to investigate seasonal migrations, or hours for daily motion patterns). The interestingness tests in Laube and Purves (2005) confirmed, for example, that in the Porcupine caribou data sample with a sampling rate δt of two weeks no interesting constancy patterns could be expected in both motion azimuth and speed. However, varying the analysis granularity may be used as an exploratory strategy in order to uncover interesting patterns at different granularities.

4.3 Recalling the research questions

The aim of this section is to examine the work presented in the publications with respect to contributions from the state of the art. Therefore the section recalls the motivating research questions, and relates the answers given in this thesis to solutions in the literature.

4.3.1 Do districts flock?

Research question

$\Longrightarrow$ *Are there generic motion patterns, i.e. equal or at least similar patterns that can be found in the tracks of MPOs modelling various motion phenomena?*

A school of fish

A linguistic excursion may give first hints about the existence of generic motion patterns. The American Heritage Dictionary of the English Language (fourth edition) gives no less than 38 terms for flock. In general, *flock* stands for (1) a group of animals that live, travel, or feed together, (2) a group of people under the leadership of one person, and (3) a large crowd or number. Some examples illustrate the diversity: *Flock* of goats or sheep, *pride* of lions, *school* of fish, *gam* of whales, *gaggle* of geese, *murder* of crows, or *swarm* of insects. Some terms have human connotations, for example *flock* and *herd* for people of a religious movement ("But he led his own people like a flock of sheep, guiding them safely through the wilderness", Psalms, 78:44-55) or *gang* and *pack* for people "engaged in criminal pursuit" (e.g. Frank Sinatra, Dean Martin and Sammy Davis Jr. in the movie "the rat pack"). Isn't it an indication of the generic nature of a phenomenon that different contexts developed specialised terms to name that very same phenomenon? This may not be a proof in a scientific reading for the generic nature of a pattern, but it provides strong evidence.

REMO patterns

This linguistic excursion suggests searching the literature for similar motion patterns and not for similar terms denoting them, well knowing that various terms can refer to the very same phenomenon. Hence, thinking in categories rather than in terms facilitates the literature review. This thesis proposes a categorisation of a set of generic motion patterns (Laube and Imfeld, 2002). This conceptual approach suggested searching for (a) *patterns over time* (constancy, change), (b) *patterns across objects* (concurrence, opposition, dispersion), and (c) *patterns over time and across objects* (trend-setter, independent, propagation).

REMO case studies

The conceptual phase as well as the implementation phase of this thesis included intensive work with motion data originating from various fields, such as biology, ecology, soccer scene analysis, and political science. In these case studies, the following motion patterns, for example regarding motion azimuth, could be identified and have been discussed in the publications included in this thesis:

- *Constancy*: seasonal migration of caribou (Laube and Imfeld, 2002), attacking striker in a soccer game (Laube et al., 2005), political entities persistently changing their voting behaviour towards a distinct political orientation (Laube et al., 2005).
- *Concurrence*: seasonal migration of caribou (Laube and Imfeld, 2002), off-side trap in a soccer game (Laube et al., 2005), joint value changes of institutionally and culturally relevant political entities moving in an abstract ideological space (Laube et al., 2005).
- *Change*: group of caribou jointly re-orienting towards the summer and winter habitat, respectively (Laube and Imfeld, 2002), political entities changing their voting orientation within a distinct time period (Laube and Imfeld, 2002).
- *Trend-setter*: leading player anticipating the coordinated defending backwards move of the team (Laube et al., 2005), trend-setting political entities anticipating a divergence along the axis *political left vs. political right* in the Swiss society (Laube et al., 2005).

The following paragraphs further discuss the evidence of REMO patterns in various fields touching on motion such as geography, biology, or sociology.

Patterns over time

It is easy to find very convincing situations for straight trajectories of sequences of constant speed. Mining mobile trajectories in order to design LBS, Smyth (2001) cuts peoples' lifelines into chunks showing characteristic motion properties measurable by automative navigation system sensors. Searching for a highway travel sequence with speed around 100 km/h is nothing other than looking for a speed constancy. Dykes and Mountain (2003) adopt a very similar strategy identifying what they call *episodes*, in this thesis' notion 'constancy-like' patterns of similar motion properties. Behavioural ecologists approach constancy patterns by investigating directional persistence in turning angle distributions (Turchin, 1998). For instance, some copepod species indeed show intermittently constant straight sequences in their foraging behaviour (Schmitt and Seuront, 2001). Or as another example, desert ants, after having performed a circuitous foraging journey, find reliably the most direct way straight back to their nest from a distance of up to 100 m (Knaden and Wehner, 2003, 2004).

Patterns across objects

By contrast, it is much harder to find evidence of patterns across objects, since these patterns require simultaneously tracked individuals instead of just single trajectories. However, nobody would probably object to the idea that caribou often forage and migrate in a spatio-temporally coordinated way, showing a flocking behaviour (e.g. Giraldeau and Caraco, 2000). Or, aren't we not all fascinated by pictures of schooling fish[2](Hoare et al., 2000). Besides biology, sports is another field to observe the co-ordinated behaviour of MPOs. The co-ordinated forward move of a soccer team's defender row setting up an offside trap is nothing other than a *concurrence* pattern. Whereas MPOs and spectators agree on the existence of such relative motion pattern, their instantiation may be very controversial.

Patterns over time and across objects

Leadership is a widespread phenomenon in societal organisations. Wildlife biology is again a rich application field, searching for spatio-temporally explicit, motion related patterns of leadership. Leadership behaviour in relation to dominance and reproductive status has been observed in grey wolves (Peterson et al., 2002). Dominant breeding wolves showed significant frontal leadership, i.e. leading the pack during travel. Dominant wolves also initiated pack activities more often than subordinate wolves. Such behaviour corresponds to the *trend-setter* pattern (Laube and Imfeld, 2002) and its spatially constrained extension, the *leadership* pattern (Laube et al., 2004). Furthermore, the animal behaviour research community works intensively on the general topic of group decision-making in animals, searching for evidence for groups led in their activities by some dominant individuals (e.g. Lachlan et al., 1998; Conradt and Roper, 2003; Rands et al., 2003; Seeley and Visscher, 2004). Some articles even integrate social behaviour and spatial structure (e.g. Hemelrijk, 2000). Thus,

[2]Readers sharing the author's fascination for schooling fish may enjoy the movie *Deep Blue – a natural history of the oceans*, a BBC Worldwide and Greenlight Media production, 2003, http://www.deepbluethemovie.com.

trend-setter-like motion patterns of animals taking the lead, also in a spatio-temporal notion, are intensively discussed in animal behaviour sciences.

Again, this wide range of application fields all yielding equal or similar patterns may not be considered as a proof for the generic nature of the proposed motion patterns. However it again provides strong evidence. In this sense one could indeed say that political districts flock when plotted over time in an ideological space.

4.3.2 Yet another query language?

Research question

$\Longrightarrow$ *How can we describe, formalise, and thus compare delimited motion events of individuals and groups of individuals?*

Spatio-temporal query languages

The DBMS community, especially the researchers interested in STDBMS, has introduced various approaches to formalise motion (e.g. Sistla et al., 1997; Gueting et al., 2003; Grumbach et al., 2003). However, formalisms emerging from database research normally focus on querying, that is on the retrieval of stored objects, collections of objects or their observations from a database (see section 2.5).

Formalising motion patterns as spatio-temporal relations

This thesis argues that analysis must go beyond querying and requires the production of new information and knowledge that is not directly observed in the stored data (Aronoff, 1989; Golledge, 2002). In this sense, the formalism developed for REMO GKD aims neither to develop yet another DBMS query language nor to provide a tool to design a spatio-temporal database. By contrast, the REMO formalism allows the interrelation of stored data *and* on the fly derived data in an exploratory GKD process to create value-added information about motion events (Laube et al., 2005).

Syntactic approach

The REMO formalism is designed to explore spatio-temporal relations, that is relations expressed as motion patterns in the trajectories of MPOs. Motion events are intrinsically spatio-temporal entities, relating the individual histories of single MPOs with the collective history of a group. Therefore the REMO formalism follows the syntactic pattern detection approach in order to describe and detect predefined motion patterns in the geospatial lifelines of MPOs (Laube et al., 2005). After Jain et al. (2000) *syntactic pattern recognition* adopts a hierarchical perspective where a pattern is viewed as being composed of simple sub-patterns, the so-called primitives. Complex patterns are represented in terms of interrelationships between primitives. This approach allows the composition of complex motion patterns involving both individual histories and inter-object associations. The complex pattern *trend-setter*, for example, is composed of the simple patterns *con-*

stancy and *concurrence*. The syntactic structure allows description of any arbitrarily complex pattern using a set of primitives and grammatical rules.

Thus, the REMO formalism mustn't be considered as just "yet another query language", but an as integral component of a GKD approach. The power of the REMO approach lies in the interplay of the pattern formalism, the underlying data model and data structures, as well as the data mining algorithms. The latter two are discussed in the following sections.

4.3.3 About the peculiarities of Lynetta, Meilen, and Zinédine

Research question

$\Longrightarrow$ *How do we best model MPOs in order to detect motion patterns?*

Motion is a very complex phenomenon addressed in many different research and application fields, hence there surely is not a single ideal way to model MPOs. The choice of the conceptual data model for MPOs depends in fact on three factors:

1. The characteristics of the investigated motion,
2. the structure of the available motion data, and
3. the question to be answered

Case studies

The case studies used in this thesis involved animal migration (e.g. the Caribou female Lynetta), motion of abstract data points in an ideological space (e.g. the Swiss District Meilen in the Canton of Zurich), and soccer scene analysis (e.g Zinédine Zidane[3]). These diverse phenomena expose a wide variety of motion characteristics. The Porcupine Caribou data (Laube and Imfeld, 2002; Laube and Purves, 2005) provided irregularly sampled geospatial lifelines with a high sinuosity. The synthetic motion of Swiss districts, emerging from an interpolating factor analysis procedure, exposed smooth trajectories with a high directional persistence and regular time intervals. The soccer players finally, are sampled regularly at 15 fixes/sec, showing strongly oscillating trajectories due to large-scale positional inaccuracy of the tracking technique using multiple television cameras.

REMO class design

Thus, the REMO framework is designed to analyse tracking data consisting of discrete fixes regular or irregular sampling rates, and must furthermore allow regular time stamps at multiple granularities in the analysis process (Laube and Imfeld, 2002). The proposed object-oriented class design underlying the REMO GKD approach adopts a hybrid strategy, strictly separating the MPO

[3]Zidane is of course not a member of the investigated Japanese student team.

MODELLING DOMAIN and the ANALYSIS DOMAIN (Figure A.1, page 102) (Laube et al., 2005). This twofold strategy allows use of a discrete model best suited for data capture and database management, and allows maximal flexibility in calculating derivatives for the analysis process. This strategy is best illustrated in the *detached attribute functions* encapsulating the calculation of motion properties in the classes of the ANALYSIS DOMAIN (Figure 3 in Laube et al. (2005)). The detached attribute functions allow derivation of motion attributes adopting various perceptions of motion, reacting to differing characteristics of the motion behaviour (Laube et al., 2005).

Influence of research question

For analysing directional autocorrelation, a data model featuring frequent discrete steps is more suited than a MOD data model occasionally updating the MPO's direction. If, in contrast, the motion property of interest are abrupt direction changes, it would not be adequate to use a data model based on curves smoothing raw trajectories. However, which data model to choose, must in every case be decided with respect to the research question. However, the ANALYSIS DOMAIN features several implemented algorithms for smoothing and resampling lifelines as a pre-processing step. Such resampling is, for instance, needed to eliminate noise emerging from inaccurate sampling. The lifelines of the tracked soccer players have been resampled to 1 fix/sec to eliminate short term positional noise when analysing azimuth patterns (Laube et al., 2005).

4.3.4 Limits of flying through the aquarium

Research question

$\Longrightarrow$ *Can we automatically detect motion patterns in voluminous tracking data of groups of MPOs?*

Limitations of fly-throughs

The exploratory power of 2-D and especially 3-D maps, cartograms and similar visualisations is limited due to symbol overlap resulting from plotting the trajectories of many objects over long time periods. However smart exploratory visualisation techniques such as 'brushing', 'interactive perspective selection', or even 'fly-throughs' may be, finding patterns in, for example, more than 100'000 activity paths pushes the analysts' visual ability beyond its limit (Kwan, 2000, p. 200). Thus, it is not surprising that most examples in the literature adopting the space-time-aquarium only handle single objects or very limited numbers of MPOs (e.g. Kraak and Koussoulakou, 2004). To quote Forer (1998, p. 175): "In short, the methodology [i.e. the 3-D space-time geometry] made little progress in moving from one with extremely elegant conceptual approach to one with applications."

Furthermore, the suitability of descriptive statistics to detect motion patterns is limited since merely collapsing all available data into a single set of descriptive measures makes it impossible to detect inter-object relations and spatially or temporally delimited motion patterns, while mapping series of snapshots each representing a static map view of the world in motion, may exactly miss the most interesting events not caught in one of the arbitrary snapshots (Chrisman, 1998). In short, visualisation and exploratory analysis of tracking data proved to be limited to short and simple tracks of small groups of MPOs (Kwan, 2000).

Twofold data mining strategy

However, to overcome the limitations of visualisation and statistics based exploration, the REMO GKD approach is based on a twofold data mining concept for automatic detection of motion patterns.

1. In a first step the raw geospatial lifelines are transformed into a structured analysis matrix synchronising the individual's sequences of motion attributes. This first step *reduces the complexity* of the motion data.
2. *Pattern matching algorithms* are subsequently used to search for instances of the predefined motion patterns in the analysis matrix. After having enriched the motion information in the form of an analysis matrix, the pattern matching algorithms are relatively simple, similar to pattern matching in strings.

"Reduce the complexity"

Complexity reduction is a widespread process in KDD. Assuming that for exploring motion two spatial, a temporal, and possibly many more attribute dimensions are just too many to grasp, reducing the complexity to fewer dimensions facilitates visual exploration or allows the application of automatic data mining algorithms. With the time plot family Imfeld (2000) presents an application of this idea, reducing the spatial characteristics of trajectories to one dimensional attributes (e.g. distance between consecutive fixes) and plotting them against time. Thus, what was spatially overlapping is rearranged side by side, strung in a row, simplifying insights. The REMO GDK adopts a very similar idea to construct the REMO analysis matrix.

REMO matrix structure

The construction of the analysis matrix plotting motion attributes over MPOs drastically reduces the two (or even three) spatial dimensions and the many optional attribute dimensions of motion data to a 2.5 dimensional matrix. The matrix is furthermore ordered along the time axis, aligning contemporaneous events in order to facilitate the identification of collective group behaviour. This idea is very similar to what Kwan (2000) calls standardisation of trajectories in the space-time aquarium. She shifts the trajectories until all 'home-work' axes coincide. In other words, she rearranges the trajectories by placing together alike motion characteristics in order to expose outliers. Using a metaphor, reducing the complexity is very similar to what we do in our everyday live having lost track of the papers on our desk: we tidy up.

REMO pattern matching algorithms

Imfeld's dimension-reduced time plots as well as Kwan's tidied up trajectories both finally depend on a visual exploration for identification of motion patterns. REMO GKD goes one step further and combines the reduction of complexity with automatic data mining algorithms working on the simplified and restructured data. Decomposing a REMO matrix into its rows (motion attribute arrays) and columns (time-slices) allows use of derivatives of classical string pattern matching algorithms such as *Brute-Force Pattern Matching* (BFPM) or *Knuth-Morris-Pratt* (KMP) (Knuth et al., 1977).

Analogy to pattern matching strings

Brute-force pattern matching simply tests all the possible placements of P relative to T in the worst case in $O(nm)$ running time (Goodrich and Tamassia, 1998). KMP uses a failure function f for the pattern string P which encodes repeated substrings inside the pattern itself to reuse previous comparisons and thus avoid unnecessary comparisons. It achieves in the worst case a running time of $O(n+m)$ (Goodrich and Tamassia, 1998). The performance of the application prototype does not show worst case behaviour and thus did not encounter significant performance problems with the above mentioned algorithms and with the case study data used so far. With larger data volumes there may a requirement to use and develop more sophisticated pattern matching algorithms.

4.3.5 On the tracks of Waldo Tobler

Research question

$\Longrightarrow$ *How can we identify interrelated or interacting individuals (subgroups) in larger groups of MPOs?*

Living in groups

Animals live in groups to reduce predation risk, to gain foraging benefits, to find a mate, to conserve heat and water, or in the case of some fish species even to reduce the energetic costs of movement (Krause and Ruxton, 2002). One way to investigate grouping is to ask who groups with whom, where and when. Such information is coded in the geospatial lifelines of the animals, their 'grounded behaviour'.

Extending TFL

One key factor that governs the evolution and maintenance of grouping behaviour throughout the animal kingdom is spatio-temporal proximity. Thus, investigating group interaction in the tradition of spatial analysis, suggests a spatio-temporal extension of Waldo Tobler's first law of geography (TFL). According to this principle "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). In order to investigate the spatio-temporal association of group movement, nearness as a concept has to be extended to include both space and time (Miller, 2004). Knowing, full well that spatio-temporal nearness is not the only aspect that interrelates individuals, it is still a nice starting point to identify subgroups in large datasets.

Nearness in attributes

What is nearby may be interrelated. 'Near' may not only be near in space or in space-time but also near in the attributes describing motion, that is showing the same or at least similar attributes. Thus, identifying subgroups in larger groups of MPOs corresponds to the data mining task segmentation, that is clustering and classification (Fayyad et al., 1996). Thus, adopting well known clustering and classification algorithms for spatio-temporal instead of static spatial data is a promising option to identify interacting subgroups.

Nearness in REMO GKD

The REMO GKD uses nearness in a classical spatial reading (near in real space) *and* denoting to similar attributes (near in attribute space). The spatially constrained motion patterns proposed in Laube et al. (2004) directly integrate spatial constraints. The patterns *flock* and *leadership* describe a (sub-)set of the MPO group performing a group event within a spatial range and for a certain time interval. The REMO patterns across objects (*concurrence, opposition, dispersion*), in contrast denote a (sub-)set of the MPO group unified by equal or at least similar attributes describing their motion, that is nearness in the aspatial meaning (Laube and Imfeld, 2002; Laube et al., 2005).

4.3.6 Moving beyond the snapshot

Research question

$\Longrightarrow$ *Can we identify motion patterns that are intrinsically dynamic, that is do not express any pattern in either space or time alone?*

Traditional geographic analysis investigates heterogeneity in static space alone, trying to discover associations, aggregations, and topological relationships of spatial entities (Golledge, 2002; Miller and Wentz, 2003; O'Sullivan and Unwin, 2003). On the other hand, time series or trend analysis methods adopted, for instance, in the fields of ecology or economics, are developed to find patterns in time alone and thus neglect the spatial dimension (e.g. Bjørnstad and Grenfell, 2001).

However, when it comes to understanding motion phenomena, both dimensions must be integrated (Massey, 1999). Consider as a first example the investigation of mobile phone tracks during a crime investigation in order to find suspects of a past hijacking. Assuming that the victim and the hijackers jointly tried to escape the police during the chase, they leave identical or at least very similar motion patterns in space-time, patterns to be identified using a suitable data mining algorithm (Summers, 2003). Or consider as a second example the motion process performed by a set of animals converging to a water hole in the savannah during an intensifying drought. In the initial phase of this process no spatial cluster can yet be seen in the distribution of the MPOs. The intrinsically spatio-temporal nature of a convergence behaviour can only be detected by a repeated inspection of the spatial aggregation over time.

Adding spatial constraints to REMO patterns

Laube et al. (2004) present a set of motion patterns that intrinsically span space and time. The proposed pattern *flock* could be used for the crime investigation. It describes the incident of n MPOs showing the same motion attributes over a distinct temporal interval (e.g. same speed on the highway, concurrently slowing down at the gas station, thereafter concurrently speeding up again, and so forth), and are all time positioned within a constraining spatial rectangle, circle, or ellipse (Figure 2, page 207, in Laube et al., 2004). The adopted key concept is the combination of a pattern found in the relative motion of the pattern building MPOs (in this case the REMO pattern *concurrence*) with a spatial condition constraining the extent of the involved fixes. The patterns *track* and *leadership* are the spatially constrained extensions of the REMO patterns *constancy* and *trend-setter*.

Dynamic application of static spatial methods

The process pattern *convergence* is the pattern of choice for the wildlife example. The convergence pattern describes a set of m MPOs at interval i that share motion azimuth vectors intersecting within a range R of radius r (Figure 3, page 209, in Laube et al., 2004). The key concept here is the dynamic application of a traditional static concept from computational geometry. The static concept is the identification of the intersection area of directed vectors. Through its repeated application using a temporal window moving along the MPO's trajectories, the static concept is used in a dynamic way. If the MPOs actually must meet, the extended pattern *encounter* adds an extrapolated motion speed to the vectors in order to predict times of arrival in the target area.

4.3.7 The flexibility of constrained caribou

$\Longrightarrow$ *How can we evaluate the interestingness of motion patterns?*

Research question

A crucial question in pattern detection in KDD is the estimation of the interestingness of the mined patterns. If a pattern is defined as an anomalous deviation from the 'expected', then the challenge lies in defining the expected. This thesis proposed an approach to evaluate the interestingness of motion patterns by first estimating the 'expected' by generating populations of constrained random walks using Monte Carlo Simulations and, second, comparing the number of patterns found in the simulated tracks with the number found in the observation data (Laube and Purves, 2005). The basic assumption of this approach is that interesting and thus meaningful patterns have a higher ratio of occurrence to those appearing in the simulated trajectories.

Unexpectedness

Silberschatz and Tuzhilin (1996) propose unexpectedness as a measure of interestingness of patterns. They argue that patterns are interesting because they contradict our expectations, which are in turn given by our system of beliefs. One option to capture the beliefs is to formulate a statistical hypothesis. The degree of belief is then defined as a significance level for which the statistic for a certain test is on the borderline of acceptance of the hypothesis. The main drawback of this statistical approach to indicate unexpectedness is that not every belief can be formulated as a testable hypothesis (Silberschatz and Tuzhilin, 1996). For this reason the approach introduced in Laube and Purves (2005) replaced the statistical hypothesis with Monte Carlo simulations, choosing the approach of numerical experiments to learn about the expected from the stochastic properties of the simulations. Hence, in order to estimate a pattern's interestingness this thesis proposes to quantify how often a supposedly unexpected pattern occurs within the dataset and, better the likelihood of such an unexpected pattern occurring.

Synthetic motion data

Hence, the quality of the evaluation approach adopted in this thesis depends on how realistically the lifelines are synthesised. The use of the constrained random walk is a straightforward yet appropriate choice. Including step length and direction change angle distributions this approach is based on two features describing trajectories that are considered crucial in behavioural ecology (Turchin, 1998). However, constrained random walks do not include *in path auto-correlation* that may be a characteristic for certain motion processes. Certain animals, for example, may always turn East after heading North, or may express whole sequences with only very short or only particularly long steps. Such patterns could be found in the tracks of animals performing seasonal migrations, showing trajectories expressing in-path auto-correlation with respect to migratory versus sedentary intervals (Bergman et al., 2000). As an alternative to the constrained random walk so-called *transition matrices* of *Markov Chains* could be used, that explicitly allow considering in-path auto-correlation in the form of direction change matrices (Jonsen et al., 2003; Jones and Smith, 2001). Furthermore, the use of constrained random walks does not allow integration of any environmental influence on the motion. Several authors demonstrated the use of use of cellular automatons or even autonomous software agents sensing and reacting on their environment to simulate motion (e.g. Beecham and Farnsworth, 1998; Batty and Jiang, 1999). However, the aim of the evaluation study was not to adopt the most sophisticated way to simulate motion, but rather to investigate feasibility of numerical experiments to evaluate the pattern detection process.

The factors influencing REMO pattern matching...

From the numerical experiments performed in Laube and Purves (2005) several lessons could be learned estimating the interstingness of the REMO patterns. First, a set of factors influences the results of the pattern detection process: (a) The properties given by the data (temporal granularity). (b) The granularity of the attribute classification (e.g. two classes with 180° azimuth intervals, four classes with 90° intervals, eight classes with 45° intervals, ...). (c) The pattern extent p (pattern length with patterns over time, pattern width for patterns across objects).

...are interrelated

A second important lesson learned is that these factors are interrelated. The numerical experiments for concurrence affirm, for instance, the straightforward hypothesis about the relation between attribute classification and pattern extent (pattern length/width). The number of MPOs divided by the number of attribute classes gives a lower threshold of pattern interestingness. Having 185 MPOs and 4 classes (8 classes), concurrence patterns having a width below 46 (below 23) are expected simply by randomly sampling motion azimuths from the 4 classes (8 classes). One way to normalise these influences is to measure the success of a different pattern detection configurations by calculating the ratio of found patterns to the patterns possible. If the number of found patterns increases due to a certain configuration of attribute granularity and pattern extent, then also the number of possible patterns increases, hence, the ratio *found / possible* remains a reliable means to estimate the success of the pattern matching.

Suitable pattern matching configurations

Overall, the evaluation method introduced in Laube and Purves (2005) is a simple yet feasible way to estimate the interestingness of patterns derived from REMO GDK. The method can furthermore be used to examine useful configurations of pattern matching sessions (i.e. attribute granularities, pattern lengths), which may not be obvious. Therefore it gives REMO GDK a more objective dimension, since the composition of the searched patterns depends not solely on the (potentially biased) knowledge of the expert user.

4.4 Summary

This chapter brought together the individual contributions of the papers. Evaluating the REMO GDK approach, the chapter first critically balanced its main strengths and weaknesses. The second part discussed applications of and limitations for the REMO GDK approach, putting an emphasis on availability and suitability of data, the role of absolute space within the REMO framework, and granularity issues. Recalling the research questions the developed methods were discussed in the light of related approaches from the latest literature. The next chapter will conclude this thesis collecting its main contributions, the gained insights and providing an outlook.

Chapter 5

Conclusions

> *"Traditional spatial analytical methods were developed in an era when data collection was expensive and computational power was weak. The increasing volume and diverse nature of digital geographic data easily overwhelm mainstream spatial analytical techniques that are oriented towards teasing scarce information from small and homogeneous datasets."*
>
> Miller and Han (2001, p. 3)

This thesis presented research in methodological GIScience. The overall goal was to develop methods for analysing the motion of points. The thesis addressed the following four research objectives in detail:

Objective 1: *This thesis shall identify, characterise, and categorise a set of basic motion patterns in the lifelines of groups of MPOs, with motion patterns being a predefined formalised search template of motion attributes such as speed, change of speed, or motion azimuth.*

Objective 2: *This thesis shall develop a knowledge discovery approach integrating techniques to formally describe motion patterns and algorithms allowing detection of these patterns in case study data.*

Objective 3: *This thesis shall develop means to describe and algorithmically detect forms of temporally coordinated and spatially directed group motion patterns in trajectories of MPOs, such as flocking, converging or diverging.*

Objective 4: *This thesis shall develop statistical measures and adopt Monte Carlo techniques to quantify the relevance and interestingness of the proposed motion patterns, to distinguish random noise from information in the knowledge discovery process.*

Critically recalling the research objectives, this concluding chapter lists the main contributions and insights, before suggesting possible research avenues in the outlook.

5.1 Main contributions

This thesis provided the following main contributions with respect to the research objectives:

Identification and characterisation of a set of generic motion patterns

Objective 1

This thesis introduced a family of motion patterns based on the concept of *relative motion* (REMO). The term relative motion denotes the interrelation of motion attributes of different moving point objects (MPO) over space and over time. This thesis made use of the motion attributes speed, change of speed, motion azimuth, and sinuosity, putting a strong emphasis on motion azimuth. A first set of REMO patterns describes motion events explicitly excluding the absolute position of the MPO fixes: *constancy, change, concurrence, opposition, dispersion, trend-setter, independent*, and *propagation* (Laube and Imfeld, 2002; Laube et al., 2005). A second set of motion patterns includes spatial constraints regarding the absolute positions of the MPO: *track, flock*, and *leadership* (Laube et al., 2004). A third set describes aggregation and disaggregation patterns: *convergence, divergence, encounter*, and *breakup* (Laube et al., 2004).

Objective 3

Objective 2

Pattern description formalism In parallel to the family of motion patterns this thesis proposed a formalism allowing the description of the patterns in a simple, precise, and compact and way (Laube et al., 2005). The formalism is based on an extension of the commonly used regular expressions and on mathematical logic.

Development of a geographic knowledge discovery (GKD) approach for exploring the motion of MPOs The core element of this thesis was the development of the REMO geographic knowledge discovery (GKD) approach to identify generic motion patterns in geospatial lifelines. The approach places a special focus on the following crucial steps of knowledge discovery in databases (KDD):

Objectives 2 and 3

- *Data reduction and projection:* The transformation of the lifeline data into the REMO analysis matrix featuring motion attributes (such as motion azimuth, speed, change of speed, and sinuosity) allows comparison of the object's motion across objects and over time (Laube and Imfeld, 2002).

- *Exploratory analysis, model selection, and hypothesis selection:* REMO GKD adopts a syntactic pattern detection approach. Thus, REMO GKD provides a pattern description formalism to compose patterns that can subsequently be searched for in the geospatial lifelines of MPOs. The numerical experiments using Monte Carlo simulations of constrained random walks introduced in Laube and Purves (2005) can be used to focus the search for meaningful patterns.
- *Data mining:* A set of data mining algorithms facilitate the automatic search for motion patterns within large datasets of geospatial lifelines. The algorithms adopt concepts from pattern matching in strings (Laube et al., 2005) and from spatial data handling and computational geometry, respectively (Laube et al., 2004).
- *Interpreting the mined patterns:* Interactive visualisation of the mined patterns linked to the object's motion in a MPO viewer allows interpretation of the mined patterns in order to suggest avenues for further research (Laube et al., 2005). The pattern interstingness tests developed in Laube and Purves (2005) finally help to identify the interesting patterns.

Object-oriented class design for GKD in geospatial lifelines The object-based nature of MPOs, events and instances of patterns is evident, – the REMO world is an object-oriented world. The REMO GKD approach has been implemented in an object-oriented prototype in order to test and improve concepts, data structures, and algorithms. An object-oriented class design is at the core of the implementation, strictly separating the handling of the original tracking data and the data mining process. The MPO MODELLING DOMAIN features data structures to store and manipulate tracking data conserving the original sampling granularity (`MPO, Fix, Point`). Functions allowing derivation motion attributes at user defined analysis granularities for the data mining process are encapsulated in the classes of the ANALYSIS DOMAIN (`REMOMatrix, REMOPattern` (e.g. `constancy`)). Separated, as they are, from the raw data handling, these functions are called *detached attribute functions.* Objectives 2 and 3

Categorisation of MPOs One emerging task facing the massive datasets covering the motion of large numbers of MPOs is the categorisation of objects showing similar motion. Such categori- Objective 3 sations may be used to allocate specific profiles to the distinct groups, for instance, to distinguish migratory from sedentary caribou herds (Bergman et al., 2000), or to aggregate tourist movements for planning issues (Forer and Simmons, 2000; Forer et al., 2004). Static clustering performed on complete trajectories fails since categorisations of MPOs typically do not persist during the whole observation period but only over a certain period of time. The patterns *concurrence*, *flock*, *convergence*, or *divergence* address categorisation of MPOs with respect to similar motion properties, explicitly allowing for the fact that this similarity emerges only temporarily. The concepts and methods presented in this thesis are, according the author's current knowledge, among the first approaches addressing the emerging problem of categorising MPOs.

Objective 3

Aggregation in space and time The aggregation processes convergence and divergence are intrinsically dynamic, hence spatio-temporal in nature, without the possibility of existence in either space or time alone. With respect to aggregation this thesis proposes a strict separation of the process convergence and the final static cluster as its optional outcome (Laube et al., 2004). Dispersed MPOs may converge without building a detectable cluster, and *vice versa* MPOs moving in a circle satisfy the conditions of a cluster but do not converge. Methods designed to identify the static outcome of a convergence (e.g. cluster detection algorithms) are not suited to identify the dynamic process convergence. Hence, Laube et al. (2004) presented algorithms to mine trajectories for dynamic motion patterns, that is, algorithms detecting convergence/divergence patterns.

Objective 4

Evaluation framework using Monte Carlo simulations of constrained random walks In its final part the thesis presented a method to estimate the relevance of the mined patterns in REMO GDK. The method adopts the concept of indicating the interestingness of KDD patterns through their *unexpectedness*. The expected number of patterns emerging from a pattern detection session is first estimated by generating a population of constrained random walks using Monte Carlo simulations. Second, the number of patterns found in the observation data is compared to the simulations. Outliers from the expected can be assigned significance. This evaluation method can be used to examine useful configurations of pattern matching sessions. Thus, the user has an objective means at hand to compose useful and meaningful patterns.

5.2 Insights

The research undertaken in this thesis included in depth work with real motion data from various application sources. The iterative process of conceptual work and prototype implementation revealed valuable lessons about the handling of motion data and their analysis. This section summarises some crucial insights gained in the course of this work.

Dynamic phenomena require dynamic analysis concepts As argued in Laube et al. (2004) and discussed in sections 4.3.5 and 4.3.6, analysis methods developed for static spatial data are unsuitable for the analysis of dynamic phenomena, such as motion.

Fruitful integration of GIScience and KDD In recent years several authors promoted the integration of spatial data handling and data mining/KDD approaches as a powerful means to detect motion patterns in lifeline data (e.g. MacEachren et al., 1999; Miller and Han, 2001; Bédard et al., 2003; Frihida et al., 2004b). The research conducted with case studies presented in Laube and Imfeld (2002); Laube et al. (2005), and especially the numerical evaluation experiments in Laube and Purves (2005) confirm this statement.

Variability in motion phenomena This thesis modelled, investigated, and simulated motion processes from different application fields (such as wildlife biology, political science, sports, etc.), and over a huge range of spatial and temporal granularities (meters to miles, milliseconds to years). It is a major insight gained from this huge variability that motion as a phenomenon is much more complex than one might expect from the seemingly simple process of points dislocating in space!

Availability of motion data The availability of motion data is still very limited, above all in biology (perhaps somewhat surprising given the progress in positioning and tracking technology in recent years). What is especially lacking is motion data of larger numbers of concurrently tracked individuals, needed to investigate group motion behaviour. This is true for field studies of large mammals as well as for studies under laboratory condition with smaller species such as insects or fish. In many field studies groups are furthermore tracked as a whole, with only one individual collared per group, making it impossible to investigate in-group processes. Outside the field of biology the availability of motion data of groups is even worse. Be it for privacy reasons or be it for technical and organisational problems, large datasets on concurrently tracked MPOs are very difficult to access. People constantly carrying a GPS receiver are still rather technophile exceptions to the rule (Mountain and Raper, 2001b). The most accessible source of motion data is simulated data (which one can argue are what specialised data are).

Use of synthetic data for evaluation The popularity of the object-oriented paradigm and the related proliferation of agent-based simulation approaches in GIScience increases the availability of artificial motion data (Batty and Jiang, 1999). The great opportunity of artificial data is its total controllability. Every dimension of artificial data can be produced in arbitrary granularity. Artificial life forms are always visible, healthy, don't die, don't get shot, don't loose their GPS receiver, don't need privacy, and are willing the report their location at any desired time.

5.3 Outlook

Location aware devices are becoming ubiquitous and will increase our capability to collect spatio-temporal movement data by many, many orders of magnitude. Further technological advances of GPS receivers, navigation systems, PDAs, and mobile phones will inevitably lead to an increasing amount of lifeline data, suited for off-line and on-line analysis.

5.3.1 Suggested improvements

The concepts and methods proposed in this thesis can be extended in various ways, especially when relaxing its rather strong basic conditions and adopting an increasingly realistic perception of real life motion processes by allowing, for instance, imperfect trajectories, fuzzy pattern description, heterogeneous space, or by including MPO semantics. This thesis pinpointed some additional relevant issues worthy of being investigated in future research.

Uncertain and incomplete lifelines Tracking data are in many cases not perfect. Especially lifeline data emerging from biological field research suffers from uncertain or incomplete trajectories due to tracking system failures. As was discussed in section 2.3.2 on data structures for MPOs some work has already been done to handle uncertain (Moreira et al., 1999; Pfoser and Jensen, 1999; Trajcevski et al., 2004) and incomplete (Wentz et al., 2003) tracking data. However, the influence of imperfect tracking data on the results of geographic knowledge discovery and geographic data mining remains an open research issue.

Fuzzy motion patterns The REMO GKD approach uses mainly crisp concepts: it requires a crisp attribute classification and the computation of derived motion properties happens at crisp times. One could now argue that such "deterministic" patterns mismatch with the rather smooth and continuous motion processes seen in many application fields. Consider for instance the following sequence of motion azimuth values of an MPO:

$$P = S(45, 45, 45, 45, 60, 45, 45) \tag{5.1}$$

Should one now consider this rather strong persistence for one motion azimuth as being a constancy pattern? The REMO pattern description formalism allows in principle the construction of motion patterns including uncertainty and attribute ranges (Laube et al., 2005). However, the influence of fuzzy patterns on the process of geographic knowledge discovery has not yet been investigated in detail.

Motion patterns in heterogeneous space In the REMO GKD approach space is featureless and homogeneous, not constraining or affecting the motion of the MPOs. However, for many application fields space is heterogeneous, heavily influencing the motion process. Consider for example individual-based models in behavioural ecology investigating habitat preferences. So, it is one obvious option for future research to identify and characterise a set of motion patterns emerging from motion in heterogeneous space.

Include MPO semantics The work presented in this thesis has focused on the purely geometric aspects of geospatial lifelines. This approach explicitly excluded the semantics of the investigated phenomenon, that is it excluded any attribute information about the moving entities and the circumstances and environment they were moving in. Excluding the semantics was a valid approach for investigating motion with a geometric focus on points/agents moving through space. However, real life motion phenomena are usually linked to the host geography and to activity states of the individual, and do not disclose these complex interrelations in the geometry of their lifelines alone. Understanding and potentially predicting the motion of people, animals, or other agents requires the integration of the geometric properties of their motion with semantic information describing the moving entities as well as the environment harbouring the motion. For example, a social scientist analysing peoples' motion will likely see value in investigating their cultural background, their socio-economic status, their purpose of travel, or their means of transport. Wildlife biologists may want to incorporate sex, age, or the physiology of moving animals. Furthermore, any assumption of agents moving in a featureless, homogeneous space does not hold for the complex motion of intelligent agents. Much greater insights can be gained by working with (x, y, t, a) data (Forer, 2002), where a stands for the attributes of the MPOs involved. Future research must thus develop conceptual approaches and analytical tools that explore the geometric *and* the semantic properties of motion.

Multi-dimensional patterns The motion patterns proposed with the REMO GKD approach only considered one motion property at a time, either speed, change of speed, motion azimuth, or sinuosity. However, some motion processes may express characteristics in more than just one motion dimension. A foraging behaviour may, for instance, be characterised by a low speed and a high sinuosity at the same time. Similarly, investigating seasonal migration one might expect high speed values only in two directions linking the summer and the winter habitat of a species. Multi-dimensional motion patterns are another promising direction future research could take.

Alternative strategies simulating motion The evaluation method presented in (Laube and Purves, 2005) could potentially be improved using alternative approaches to simulate trajectories. The use of Markov Chains would allow including in-path auto-correlation, such as preferred turning sequences of intervals different step lengths. The use of cellular automatons or autonomous software agents would allow one to let the MPOs interact with their environment, potentially leading to a more realistic simulation of trajectories.

5.3.2 Final remarks

In recent years GIScience moved from a data poor and computation poor period to a data rich and computation rich period. Technological advances will increase this production of individualised trajectory data by orders of magnitude. Telecommunication services or customer loyalty card systems already automatically produce amounts of data that push the analysts' capabilities beyond their limits. Compared to rather visually oriented analysis techniques such as exploratory spatial data analysis geographic knowledge discovery exhibits a high potential to cope with the emergent large volumes of tracking data.

However, only the widespread application of GKD in practice can prove its usability. Up to now the applications found are very specific to some selected fields, such as transport demand modelling. Only facing more and more approaches integrating GIScience and data mining will we better understand why *spatial* (and of course spatio-temporal) is *special* with respect to knowledge discovery – and learn more about the 'G' distinguishing GKD form KDD.

Patrick Laube, April 2005.

Appendix A

Prototype implementation

> *"For a long time it puzzled me how something so expensive, so leading edge, could be so useless, and then it occurred to me that a computer is a stupid machine with the ability to do incredibly smart things, while computer programmers are smart people with the ability to do incredibly stupid things. They are, in short, a dangerously perfect match."*
>
> *Notes from a Big Country*,
> Bill Bryson (1999, p. 352).

This appendix describes the REMO prototype implementation. The prototype has been developed to conceptualise, test, and improve the methods introduced in this thesis. The REMO application is a stand-alone Java application and hence platform independent. This appendix first illustrates the object-oriented class design building the basis for the implemented pattern detection techniques. Second most important features of the graphical user interface are introduced, putting an emphasis on functionalities that can't normally be found in static out-of-the-box geographical information systems. Third, the example of searching a trend-setting caribou anticipating the migration of the whole herd illustrates the use of the prototype for the REMO GKD process.

A.1 The REMO class design

The UML class diagram in Figure A.1 provides a generalised overview of the object-oriented design of the REMO prototype application, excluding the whole GUI. This generalised overview allows the interested reader to see in which classes the crucial functionalities of REMO GKD are encapsulated, without getting lost

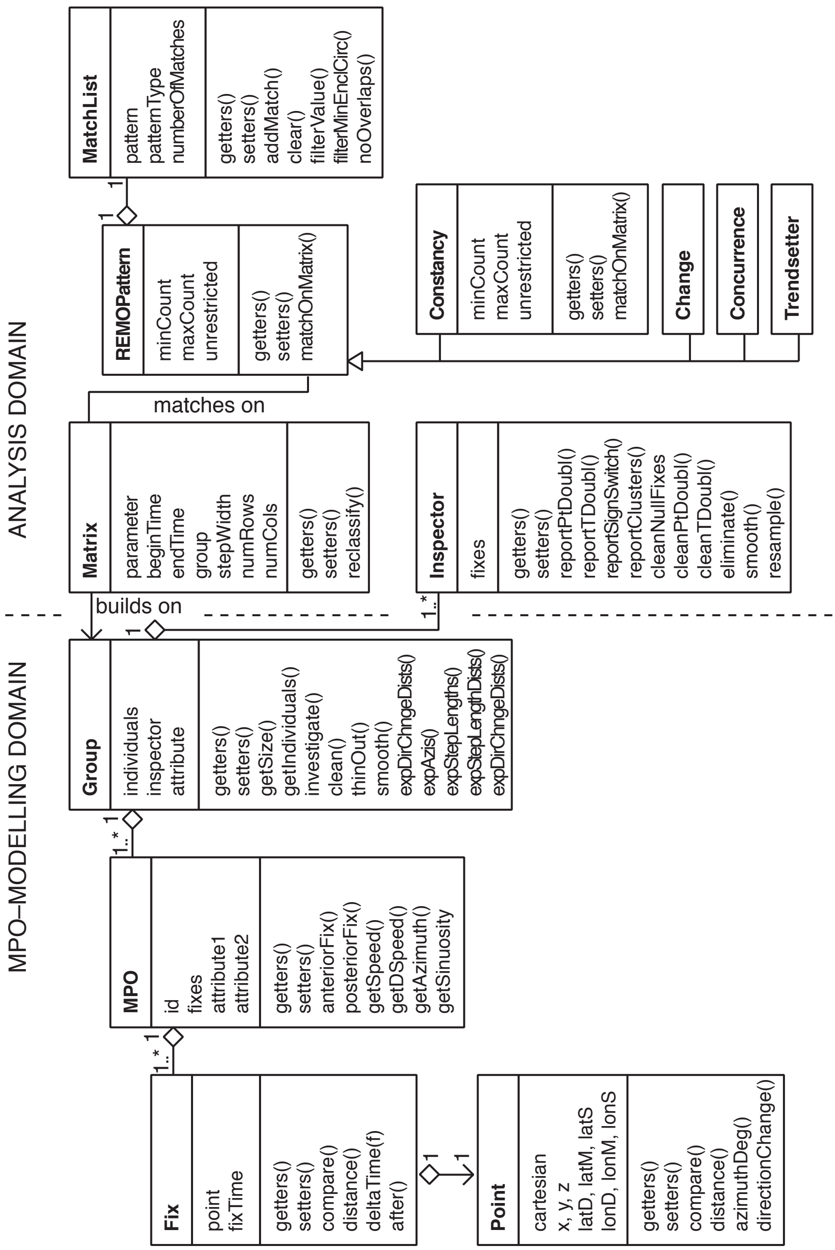


Figure A.1: UML class design of the REMO GKD prototype application.

in coding details. However, this section places a special emphasis on a set methods crucial for the REMO GKD approach.

Detached attribute functions

The derivation of MPO motion properties at arbitrary times was a core requirement for the REMO class design. The so-called *detached attribute functions* provide this functionality (Laube et al., 2005). The detached attribute functions are implemented as a wide range of polymorphic methods encapsulated in the class `MPO`[1]. The methods `MPO.getSpeed()`, `MPO.getDSpeed()`, `MPO.getAzimuth()`, and `MPO.getSinuosity()` take a query time and a temporal interval as arguments to derive the requested motion attribute from the original tracking data stored in the MPO's collection of `Fixes`.

Pattern matching

The method `REMOPattern.matchOnMatrix()`, the central pattern detection method, is encapsulated in the class `REMOPattern`, respectively in its child classes `Constancy`, `Change`, `Concurrence`, and `Trend-setter`. Thus, the `REMOPatterns` match instances of themselves on the `Matrix`, subsequently reporting the found instances to a `MatchList` object associated to every pattern detection run.

Filtering

The method `REMOPattern.matchOnMatrix()` constructs an instance of the class `MatchList`. This class is used to manage and manipulate the results of the pattern detection process. The main manipulation in this final step of the data mining process is *filtering*. Filtering for attributes using `MatchList.filterValue()` is used to select only those patterns of a specified attribute value (e.g. only azimuth `Concurrence` patterns of 45°). Filtering using spatial constraints is another way of filtering a MatchList. For example, the method `MatchList.filterMinEnclCirc()` selects only those instances of a `Concurrence` MatchList that lie within a specified minimal enclosing circle, thus building a `Flock` pattern.

A.2 Graphical User Interface

The implemented graphical user interface (GUI) allows a potential user to perform the whole REMO GDK process in an interactive and thus exploratory mode. Therefore the GUI adopts many visualisation concepts that appear to be very efficient in visualising the motion of MPOs (section 2.6), including animation in map views, colour coding for attributes, point density representations, and interactive linking between different analysis frames.

[1] This chapter uses `typewriter` referring to elements of the class design and **bold sans-serif** referring to elements of the graphical user interface.

A.2.1 Main frame

The main frame features a menu bar and a logger window listing the crucial settings and outputs of the pattern detection session. The following paragraphs describe the main menus (Figure A.2).

File menu As usual, the file menu controls file input and output. The prototype application reads simple comma delimited ascii files containing the fixes of the MPOs.

Group menu This menu features methods to manipulate a whole group of MPOs, that is, methods applied to all individuals of a group. This includes first various methods to inspect and subsequently clean the raw trajectories in order to detect and eliminate outliers, sign switches, as well as time and location doublets. The menu also features methods to thin out and smooth the raw trajectories, a useful option for large and noisy data sets. This menu allows the user furthermore to compute and export a set of statistics files, describing the motion properties of a groups' individuals. These properties include for instance step length and turning angle distributions angles characterising the trajectories of the individual MPOs. Finally the plot item launches the space-time viewer (see Figure A.3).

`group.investigate()`
`group.clean()`
`group.thinout()`
`group.smooth()`
`group.expDirChngeDists()`
`group.expStepLengthDists()`
`group.expAzimuthValues()`

Matrix menu This menu controls the construction of the REMO analysis matrix. First, a dialogue allows the specification of the general configuration of the pattern detection session. The user can set the motion parameter under study (i.e. azimuth, speed, δ-speed, or sinuosity), the step width of the matrix (i.e. the interval δt calling the detached attribute functions such as `MPO.getSpeed()`), as well as the begin time and end time of the matrix. The user then has to specify the reclassification mode by selecting a number of classes for a linear reclassification or by selecting a remap-table (rclss file). Finally the user may construct the REMO matrix and its visualisation by selecting a colour coding mode, setting up a blue-red ramp or selecting an arbitrary colour map (colour map file).

`matrix.reclassify()`

subGrouping menu This menu contains a single entry, creating a so-called *distance matrix* and some additional output files. This command launches a set of methods that are not explicitly discussed in this thesis. Those methods are used to identify subgroups of MPOs expressing similar motion behaviour.

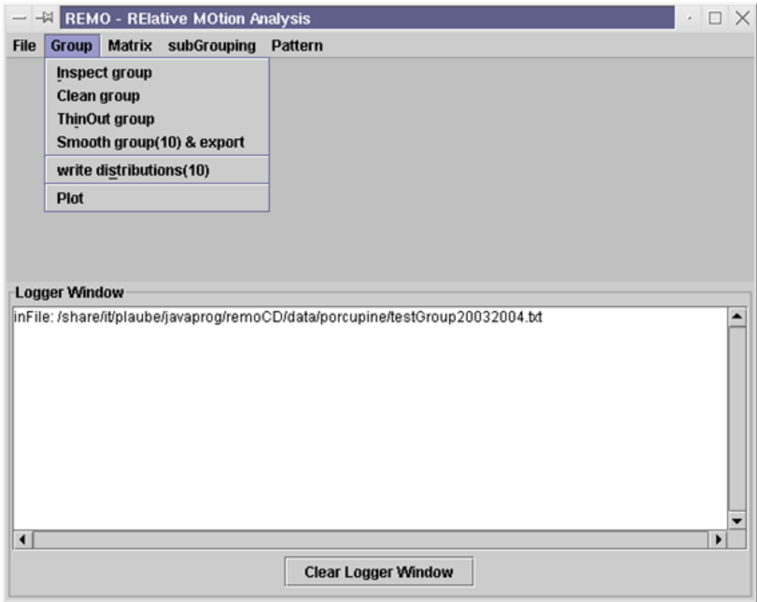


Figure A.2: The main frame of the REMO GUI.

Pattern menu The pattern menu controls the pattern detection sessions performed on the previously constructed REMO matrix. The four menu items allow the user to construct patterns that are subsequently searched on the REMO matrix. Every menu item launches a dialogue to specify the characteristics of the pattern.

`REMOPattern.matchOnMatrix()`

A.2.2 Map viewer

Decomposing trajectories

With its **map viewer** the REMO GUI features a simple yet effective tool to visualise the trajectories of a group of MPOs. The map viewer features a conventional geo-view, a toolbox providing a panning and a zooming tool and a set of display controlling elements (Figure A.3, page 106). As an exploratory tool the map view allows the decomposition of the confusing overlaps of too many fixes into smaller subsets, be it spatially or temporally delimited (Figure 2.3, page 21). Thus, the map viewer implements the concept of *space-times* and *time-spaces* introduced by Dykes and Mountain (2003). Like many other exploratory applications, the REMO prototype adopted this strategy by featuring a *bitemporal moving window*. Two simple sliders allow the user to browse through the motion data, the first controlling the current time and second the length of the plotted temporal interval. After having

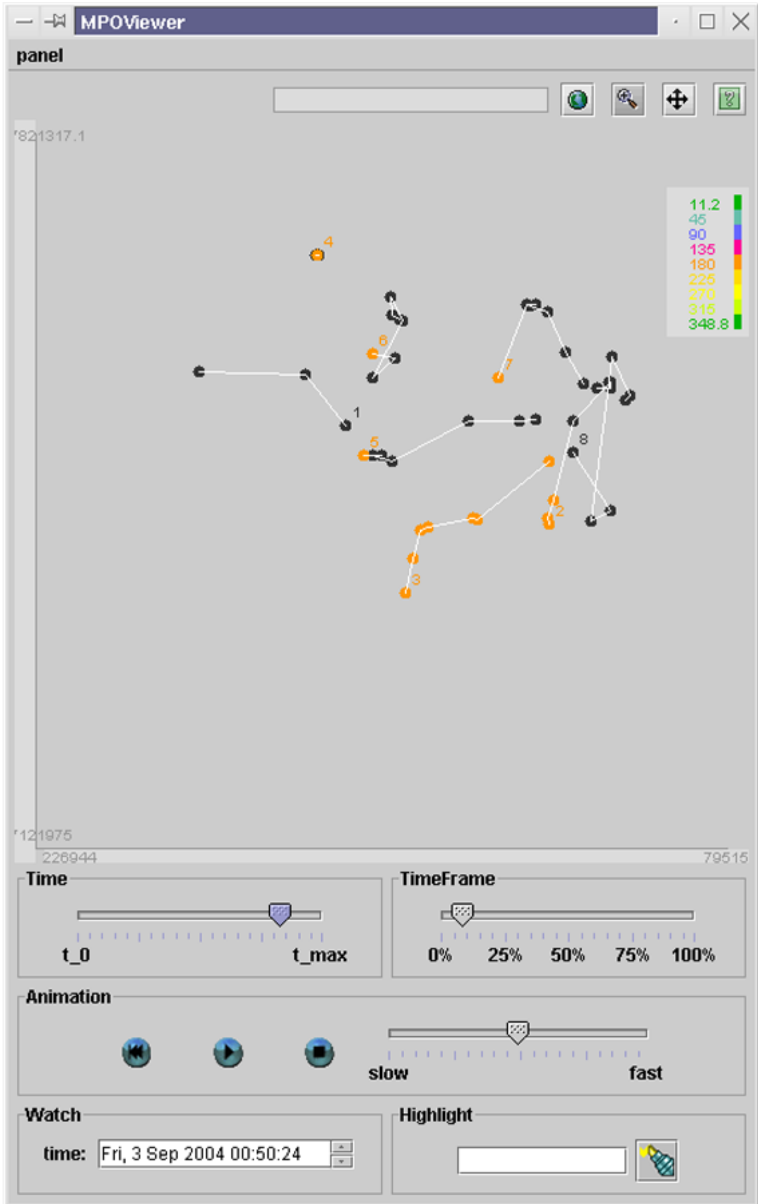


Figure A.3: The **map viewer** visualises the investigated lifelines. Two sliders control the displayed interval: one sets the current time, the other sets the length of the interval displayed. Mined patterns are highlighted in the map view for exploratory analysis. In this case the mined trend-setter patterns of Figure A.9 is highlighted.

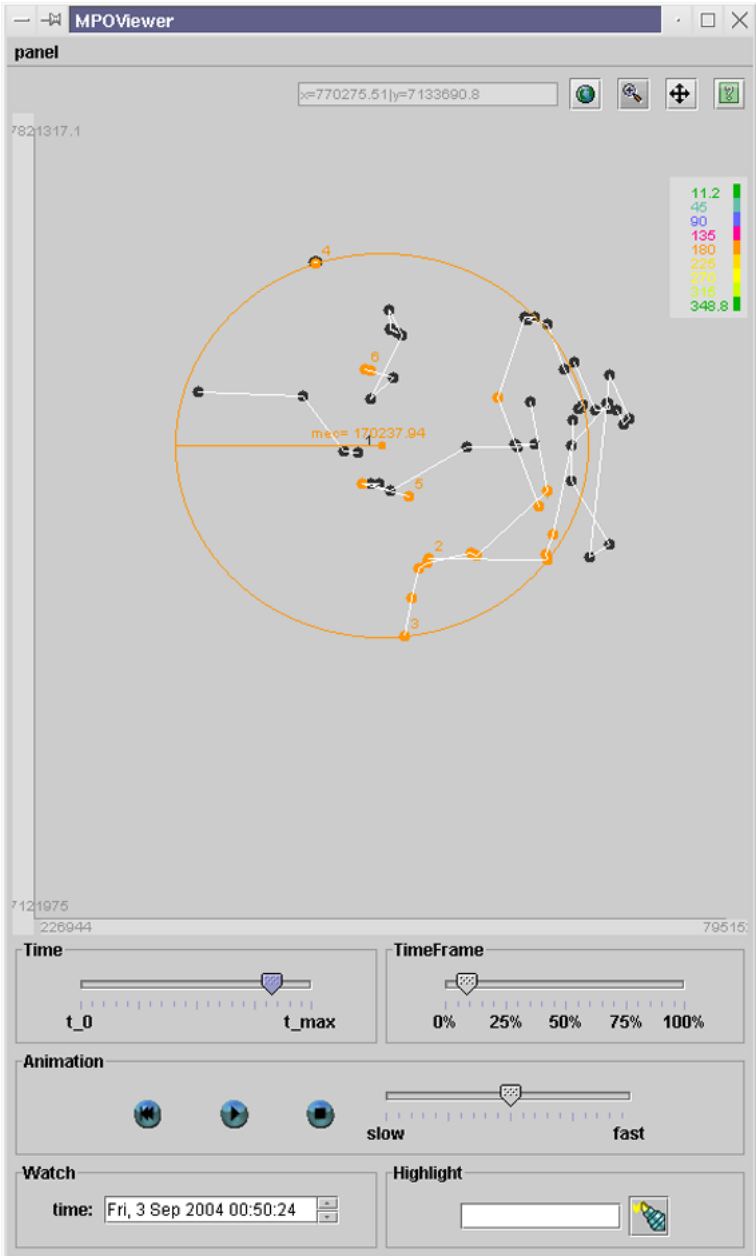


Figure A.4: Example for a spatially constrained `Trend-setter` pattern, visualised in the **map viewer**.

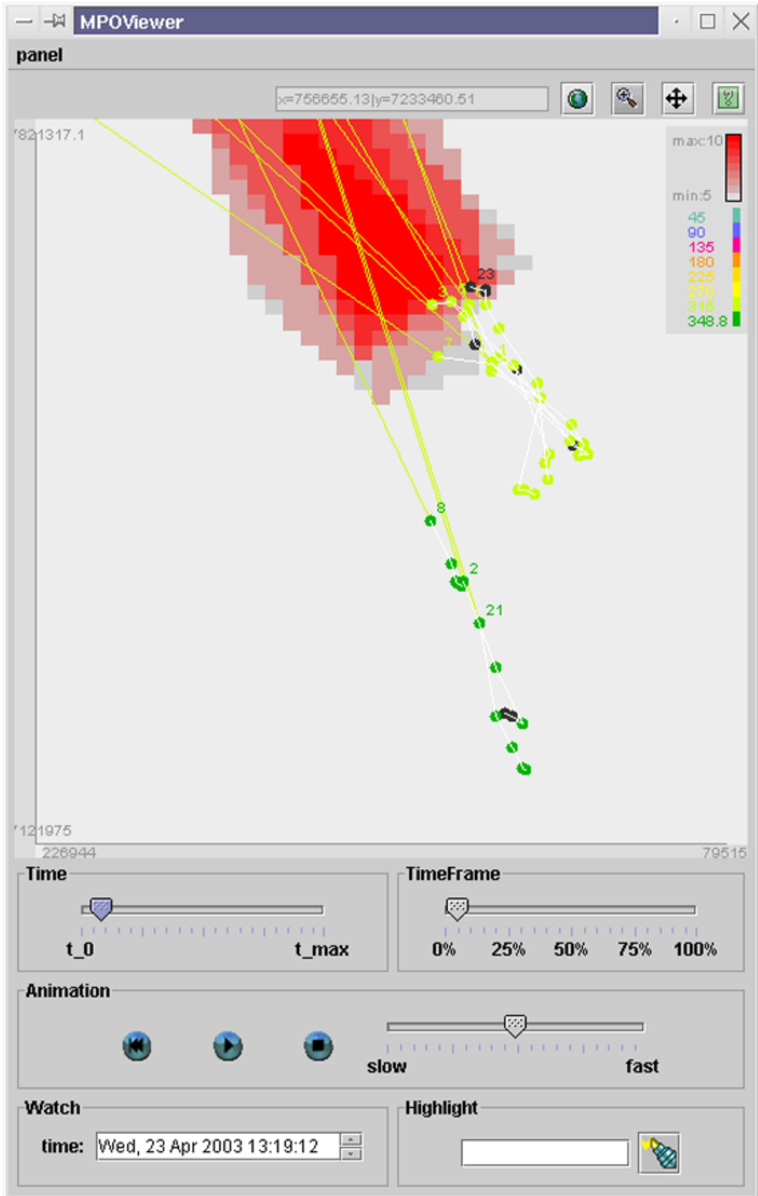


Figure A.5: The pattern `Convergence` identifies areas where many MPOs are heading for. The dynamically computed grid highlights such areas with red colour. The **heading tool** furthermore extrapolates the motion direction each MPO is currently heading for.

selected a temporal interval, the map viewer allows the animation of the motion process. Alternatively, any arbitrary point in time can be plotted using the watch.

Lifelines code for attributes

The display of the lifelines can be adapted in multiple ways. First, the colour of the lifelines can express several attributes. In the **plain** mode, all fixes are black, allowing the user to highlight mined patterns (Figure A.3, page 106). In the **azimuth** mode each fixes express the motion azimuth to the previous fix, colour coded with eight azimuth classes. The **worm** mode illustrates all fixes of the plotted time frame with a gray-red ramp expressing dynamics. And, finally, the **shuffle** mode assigns every MPO its own colour.

Heading tool

The prototype features some additional methods implemented for the spatially constrained patterns. The **heading** tool extrapolates the current motion of the MPOs. The heading is a directed ray attached at the last fix plotted and pointing to the mean direction of the currently plotted trajectory of an MPO, constrained by the plotting interval (Figure A.5).

Minimal enclosing circle

REMO patterns have a spatial extent, representing the area covered by all fixes involved in a pattern. The area including all involved fixes could be represented as the minimal enclosing rectangle, their convex polygon or, as is implemented in the prototype, as a simple minimal enclosing circle (Figure A.4, page 107). This tool helps, for example, to visualise the difference between a simple `Concurrence` pattern and its spatially constrained equivalent `Flock` (Laube et al., 2004).

Convergence grid

The concept of convergence (Laube et al., 2004) has been implemented using a grid based approach (Figure A.5, page 108). The areas where many MPOs are converging are highlighted in red, the more MPOs, the higher the red colour saturation. This area of convergence changes and dislocates over time, dynamically indicating where the MPOs are currently heading for.

A.2.3 MatchList viewer

The result of every pattern detection session is illustrated within the **matchList frame** (Figure A.9, page 112). The left top panel highlights the detected instances of the mined patterns. The patterns plotted are clickable, in order to retrieve their corresponding meta-information and to be displayed in the two logger windows at the bottom.

A.3 The REMO pattern detection process

This section exemplifies the REMO GKD process. For this example motion data is taken from the Porcupine Caribou Herd

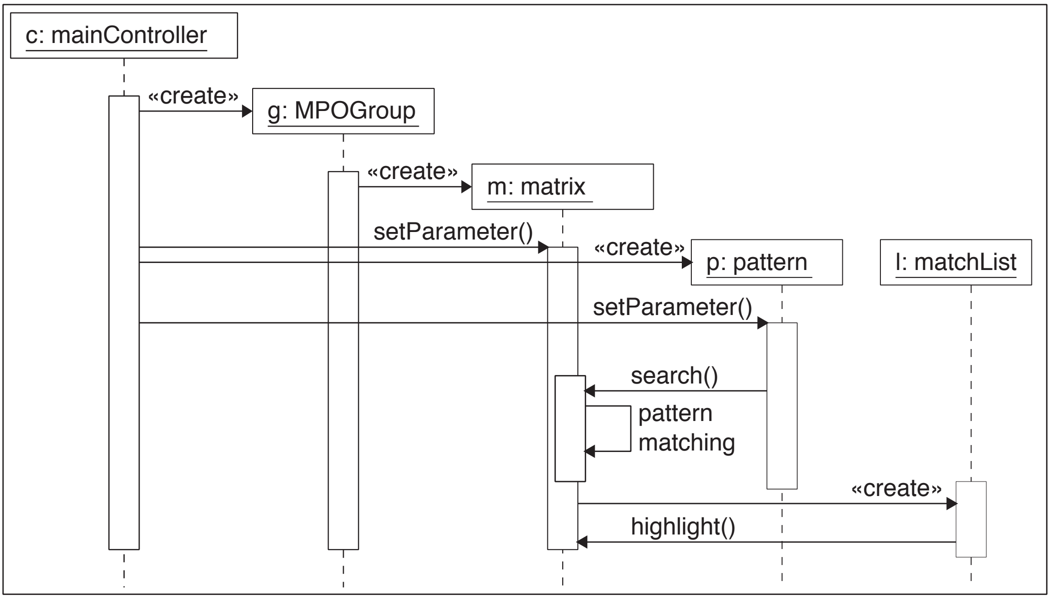


Figure A.6: UML process model of the REMO GKD process.

(PCH), referred to in (Laube and Imfeld, 2002) and Laube and Purves (2005). In short, the data sample includes ten caribou individuals, fixed every second week from March 2003 until the end of 2004 (see Figure XY in Laube and Purves (2005)). Consider for example wildlife biologists studying the seasonal migrations of the PCH. They are especially interested in individuals that lead the herd into the seasonal migration. In the terminology of this thesis, such a motion event corresponds to the pattern `Trend-setter` potentially found in the motion azimuth of the lifelines. From their expert knowledge they compose a pattern of interest, consisting of individuals that anticipate 3 time steps in advance the motion of at least 5 followers. Following Laube et al. (2005) this pattern P is formalised as

$$P = \begin{cases} S([\sharp]\{3\}) & : \quad t_0 \cdots t_2 \\ I([\sharp]\{5,\}) & : \quad t_2 \end{cases} \tag{A.1}$$

After having started the REMO prototype application, the user selects the file containing the lifelines of the caribou under study. Thereafter the application creates a `MPOGroup` (see Figure A.6 top left). After some optional pre-processing steps from the **group menu** (e.g data cleaning), the user sets the parameters of the pattern matching session (**matrix menu**), including the motion prop-

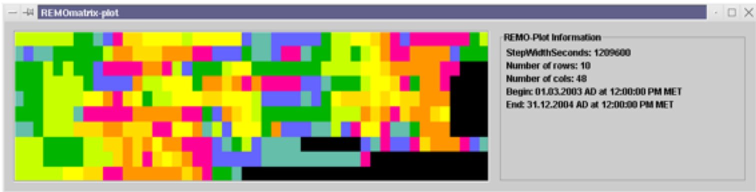


Figure A.7: Matrix frame illustrating the colour coded motion azimuth of ten female caribou from the PCH, δt is two weeks. See Figure A.3 for a legend of the colour code, black refers to noData.

erty (azimuth) and the step length of the REMO `matrix`. After confirmation, the application creates the `matrix` visualised in the **matrix frame** (see Figure A.7).

The next step consists of constructing the pattern search template. The user calls **trend-setter** from the **pattern menu** (see Figure A.8). Input fields allow the user to specify the extent of the primitives building the `Trend-setter`, that is the length 3 of the pattern `Constancy` and the width $\geq$5 of the pattern `Concurrence`. The **Match of Matrix** button launches the pattern matching algorithms, resulting in the display of the **matchList frame** (see Figure A.9).

The matches of P listed and illustrated in the **matchList frame** are now interactively linked with the **map view** (see Figure A.3), thus providing an interactive way of exploring the results of the GDK process. As the user browses through time, the `fixes` involved in an instance of a pattern are highlighted in the respective colour. In Figure A.3 the trend-setting caribou Lynetta anticipates the southward motion of six followers during fall migration.

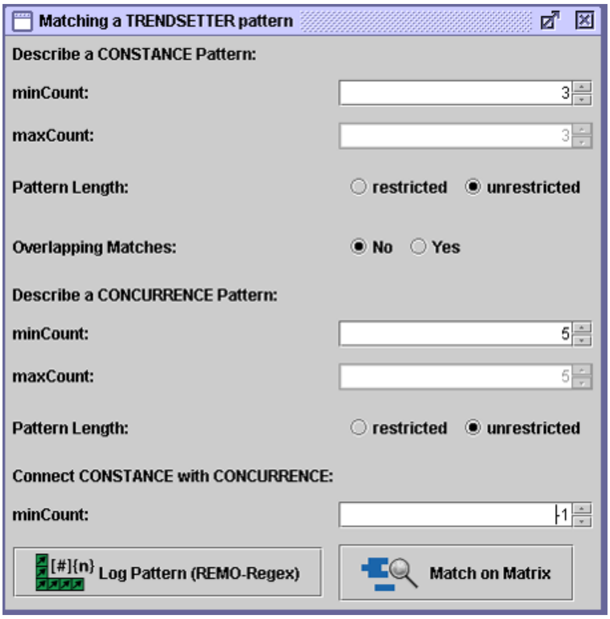


Figure A.8: Constructing the `Trend-setter` P.

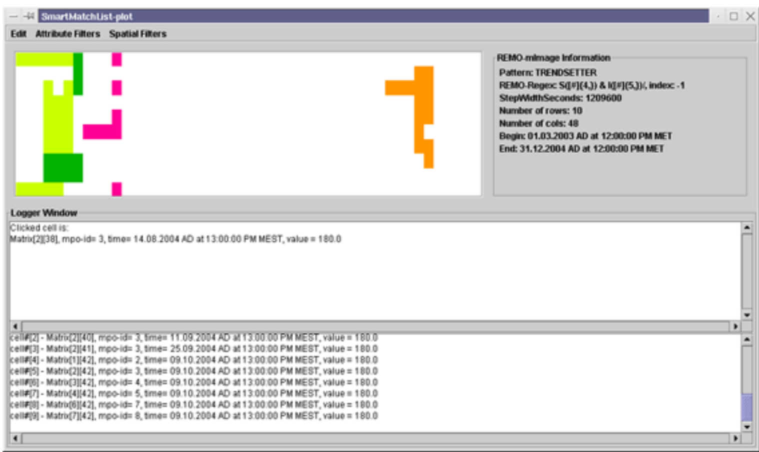


Figure A.9: The **matchList frame** collects and visualises the results of the pattern detection process. The colour coded patterns are illustrated top left and their meta-information is logged at the bottom. The patterns are clickable, jumping to the meta-information in the logger frames. The patterns represent trend-setting caribou anticipating the northward and southward seasonal migration.

Appendix B

Complete publication list

The research done in the course of this thesis resulted in a set of additional publications that have not explicitly been introduced in this volume. These additional publications are marked with a star *.

*Laube, P. (2001). A Classification of Analysis Methods for Dynamic Point Objects in Environmental (GIS). In Konecny, M., editor, *GI in Europe: Integrative, Interoperable, Interactive, Proc. of the 4th AGILE Conference, Brno, Czech Republik, April 19th - 21th, 2001*, pages 121–134.

Laube, P. and Imfeld, S. (2002). Analyzing relative motion within groups of trackable moving point objects. In Egenhofer, M. J. and Mark, D. M., editors, *Geographic Information Science*, volume 2478 of *Lecture Notes in Computer Science*, pages 132–144. Springer, Berlin-Heidelberg, DE.

*Laube, P. , and Imfeld, S. (2003). REMO-Regex: Matching Motion Patterns in Groups of Moving Point Objects. Poster Presentation, Conference on Spatial Information Theory (COSIT), September 2003, Kartause Ittingen, Switzerland.

Laube, P., Van Kreveld, M. and Imfeld, S. (2004). Finding REMO - detecting relative motion patterns in geospatial lifelines. In Fisher, P. F., editor, *Developments in Spatial Data Handling*, Proceedings of the 11th International Symposium on Spatial Data Handling, pages 201–215. Springer, Berlin-Heidelberg, DE.

*Laube, P., and Purves, R.S. (2005). Evaluation of a geographic knowledge discovery approach using random walk models. In Billen, R., Drummond, J., Forrest, D., and João, E., editors, *Proceedings of the GIS Research UK: GISRUK 2005*, pages 130–135, University of Glasgow.

Laube, P., Imfeld, S., and Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668.

Laube, P. and Purves, R.S (submitted 2005). Evaluating motion pattern techniques in spatio-temporal data. *Computer, Environment and Urban Systems*, Submitted April 2005.

Appendix C

Curriculum vitae

PATRICK OLIVIER LAUBE
born October 21st, 1972, in Kölliken, AG, Switzerland
citizen of Baldingen, AG, Switzerland.

Education

1989 – 1993	High school in Aarau (Alte Kantonsschule Aarau), concluded with "Matura" exam type "C" (preference in mathematics and sciences).
1993 – 1999	Studies in Geography, Sciences faculty, University of Zurich, minors in biology and geology (at Swiss Federal Institute of Technology Zurich, ETHZ).
1999	MSc Thesis. Lösungen zur Datenerhebung und Datenintegration in der Huftierforschung des Schweizerischen Nationalparks (Data Capture and Data Integration Solutions for Wildlife Science in the Swiss National Park), advised by Prof. Dr. K. Brassel, Dr. B. Allgöwer, Dr. A. Streilein, and R. Haller.
1999	Diploma in Geography (dipl. geogr.), University of Zurich, Switzerland.
2000 – 2005	Teaching assistant, Geographic Information Systems Division, Department of Geography, University of Zurich.
2002 – 2005	Dissertation at the Geographic Information Systems Division, Department of Geography, University of Zurich. Title of thesis "Analysing Point Motion – Spatio-Temporal Data Mining of Geospatial Lifelines", advised by Dr. Stephan Imfeld, Prof. Dr. R. Weibel, Prof. Dr. Peter F. Fisher, and Dr. Britta Allgöwer.

Bibliography

Abraham, T. and Roddick, J. F. (1999). Survey of spatio-temporal databases. *GeoInformatica*, 3(1):61–99.

Agarwal, P. K., Arge, L., and Erickson, J. (2003). Indexing moving points. *Journal of Computer and System Sciences*, 66(1):207–243.

Agarwal, P. K., Guibas, L. J., Edelsbrunner, H., Erickson, J., Isard, M., Har-Peled, S., Hershberger, J., Jensen, C., Kavraki, L., Koehl, P., Lin, M., Manocha, D., Metaxas, D., Mirtich, B., Mount, D., Muthukrishnan, S., Pai, D., Sacks, E., Snoeyink, J., Suri, S., and Wolfson, O. (2002). Algorithmic issues in modeling motion. *ACM Comput. Surv.*, 34(4):550–572.

Allen, J. F. (1984). Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154.

Andrienko, N. V. and Andrienko, G. L. (1999). Interactive maps for visual data exploration. *International Journal of Geographical Information Science*, 13(4):355–374.

Andrienko, N. V., Andrienko, G. L., and Gatalsky, P. (2000). Towards exploratory visualization of spatio-temporal data. In *3rd AGILE Conference on Geographic Information Science*, pages 137–142.

Anselin, L. (1990). What is special about spatial data? In Griffith, D., editor, *Statistics, past, present, and future*, Monograph Series, pages 63–77. Institute of Mathematical Geography, Ann Arbor, MI.

Anselin, L. (1998). Exploratory spatial data analysis in a geocomputational environment. In Longley, P. A., Brooks, S. M., McDonnell, R., and Macmillan, B., editors, *Geocomputation: a primer*, chapter 5, pages 77–94. John Wiley & Sons, New York, NY.

Aronoff, S. (1989). *Geographic Information Systems: A Management Perspective*. WDL Publications, Ottawa, CA.

Basch, J., Guibas, L. J., and Hershberger, J. (1997). Data structures for mobile data. In *Proceedings of the eighth annual ACM-SIAM symposium on discrete algorithms*, pages 747–756, Philadelphia, PA. Society for Industrial and Applied Mathematics.

Batty, M., Desyllas, J., and Duxbury, E. (2003). The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades. *International Journal of Geographical Information Science*, 17(7):673–697.

Batty, M. and Jiang, B. (1999). Multi-agent simulation: new approaches to exploring space-time dynamics within GIS. Technical report, Centre for Advanced Spatial Analysis, University College, London, UK.

Bédard, Y., Gosselin, P., Rivesta, S., Proulx, M.-J., Nadeau, M., Lebel, G., and Gagnon, M.-F. (2003). Integrating GIS components with knowledge discovery technology for environmental health decision support. *International Journal of Medical Informatics*, 70(1):79–94.

Beecham, J. A. and Farnsworth, K. D. (1998). Animal foraging from an individual perspective: an object oriented model. *Ecological Modelling*, 113(1–3):141–156.

Berger, U., Wagner, G., and Wolff, W. F. (1999). Virtual biologists observe virtual grasshoppers: an assessment of different mobility parameters for the analysis of movement patterns. *Ecological Modelling*, 115(1–2):119–127.

Bergman, C. M., Schaefer, J. A., and Luttich, S. (2000). Caribou movement as a correlated random walk. *Oecologia*, 123(3):364–374.

Berry, B. J. L. (1964). Approaches to regional analysis: A synthesis. *Annals of the Association of American Geographers*, 54(1):2–11.

Bjørnstad, O. N. and Grenfell, B. T. (2001). Noisy clockwork: Time series analysis of population fluctuations in animals. *Science*, 293(5530):638–643.

Brillinger, D. R., Preisler, H. K., Ager, A. A., and Kie, J. G. (2001a). The use of potential functions in modelling animal movement. Technical report, Department of Statistics, University of California, Berkeley.

Brillinger, D. R., Preisler, H. K., Ager, A. A., and Kie, J. G. (2004). An exploratory data analysis (eda) of the paths of moving animals. *Journal of statistical planning and inference*, 122(2):43–63.

Brillinger, D. R., Preisler, H. K., Ager, A. A., Kie, J. G., and Stewart, B. S. (2001b). Modelling movements of free-ranging animals. Technical Report 610, Department of Statistics, University of California, Berkeley, CA.

Broener, K. and Penumarthy, S. (2003). Social diffusion patterns in three-dimensional virtual worlds. *Information Visalization*, 2(3):182–198.

Bryson, B. (1999). *Notes from a Big Country.* Black Swan Books, Transworld Publishers Ltd, London, UK.

Burrough, P. A. and McDonnell, R. (1998). *Principles of Geographical Information Systems*. Oxford University Press, Oxford, UK.

Byers, J. A. (2001). Correlated random walk equations of animal dispersal resolved by simulation. *Ecology*, 82(6):1680–1690.

Cain, M. L. (1989). The analysis of angular data in ecological field studies. *Ecology*, 70(5):1540–1543.

Carter, J. and Finn, J. T. (1999). MOAB: a spatially explicit, individual-based expert system for creating animal foraging models. *Ecological Modelling*, 199(1):29–41.

Cheng, R., Kalashnikow, D. V., and Prabhakar, S. (2004). Querying imprecise data in moving object environments. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1112–1127.

Chrisman, N. R. (1998). Beyond the snapshot: Changing the approach to change, error, and process. In Egenhofer, M. and Golledge, R., editors, *Spatial and Temporal Reasoning in Geographic Information Systems*, chapter 6, pages 85–93. Oxford University Press, Oxford, UK.

Cliff, A. and Haggett, P. (2004). Time, travel and infection. *British Medical Bulletin*, 69(1):87–99.

Cliff, A., Haggett, P., and Smallman-Raynor, M. (1993). *Measles: An Historical Atlas of a Major Human Viral Disease*. Blackwells, Oxford, UK.

Conradt, L. and Roper, T. (2003). Group decision-making in animals. *Nature*, 421(6919):155–158.

Couclelis, H. (1999). Space, time, geography. In Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W., editors, *Geographical Information Systems*, volume 1, chapter 2, pages 29–38. John Wiley & Sons, New York, NY, 2nd edition.

Deutsch, D. (1998). *The Fabric of Reality*. The Penguin Press, London, UK.

DiBiase, D., MacEachren, A. M., Krygier, J. B., and Reeves, C. (1992). Animation and the role of map design in scientific visualization. *Cartography and Geographic Information Systems*, 19(4):201–214.

Dobson, J. E. and Fisher, P. F. (2003). Geoslavery. *IEEE Technology and Society Magazine*, 22(1):47–52.

Doherty, S. T. and Miller, E. J. (2000). A computerized household activity scheduling survey. *Transportation*, 27(1):75–97.

Dutton, G. (1999). Scale, sinuosity and point selection in digital line generalisation. *Cartography and Geographic Information Systems*, 26(1):33–53.

Dykes, J. A. and Mountain, D. M. (2003). Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and application. *Computational Statistics and Data Analysis*, 43(4):581–603.

Egenhofer, M. J. and Franzosa, R. D. (1991). Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2):161–174.

Erwig, M., Güting, R. H., Schneider, M., and Vazirgiannis, M. (1999). Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica*, 3(3):269–296.

Erwig, M. and Schneider, M. (2003). A visual language for the evolution of spatial relationships and its translation into a spatio-temporal calculus. *Journal of Visual Languages and Computing*, 14(2):181–211.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54.

Forer, P. (1998). Geometric approaches to the nexus of time, space, and microprocess: Implementing a practical model for mundane socio-spatial systems. In Egenhofer, M. J. and Colledge, R. G., editors, *Spatial and Temporal Reasoning in Geographic Information Systems*, chapter 13, pages 171–190. Oxford University Press, Oxford, UK.

Forer, P. (2002). Timelines, environments, and issues of risk in health: the practical algebra of x, y, t, a. In Briggs, D., Forer, P. C., Järup, L., and Stern, R., editors, *GIS for Emergency Preparedness and Health Risk Reduction*, pages 35–60. Springer, Berlin Heidelberg, DE.

Forer, P., Chen, H. F., and Zhao, J. F. (2004). Building, unpacking and visualising human flows with GIS. In *Proceedings of the GIS research UK 12th Annual Conference*, pages 334–336, Norwich, UK. University of East Anglia.

Forer, P. and Simmons, D. G. (2000). Analysing and mapping tourist flows. Technical report, Tourism, Recreation Research and Education Centre, Lincoln University, NZ.

Frank, A. U. (1998). Different types of times in GIS. In Egenhofer, M. J. and Colledge, R. G., editors, *Spatial and Temporal Reasoning in Geographic Information Systems*, chapter 3, pages 40–62. Oxford University Press, Oxford, UK.

Frank, A. U. (2001). Socio-economic units: Their life and motion. In Frank, A. U., Raper, J., and Cheylan, J.-P., editors, *Life and* *motion of socio-economic units*, volume 8 of *GISDATA*, chapter 2, pages 21–34. Taylor & Francis, London, UK.

Franke, A., Caelli, T., and Hudson, R. J. (2004). Analysis of movements and behavior of caribou (*Rangifer tarandus*) using hidden Markov models. *Ecological Modelling*, 173(4):259–270.

Franklin, S. and Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. In Muller, J., Wooldridge, M., and Jennings, N., editors, *ECAI 96: Proceedings of the Workshop on Intelligent Agents III: Agent Theories, Architectures, and Languages*, pages 21–35. Springer, Berlin Heidelberg, DE.

Frihida, A., Marceau, D. J., and Theriault, M. (2004a). Development of a process-based model for dynamic interaction in spatio-temporal GIS. *GeoInformatica*, 82(3):211–235.

Frihida, A., Marceau, D. J., and Theriault, M. (2004b). Extracting and visualizing individual space-time paths: An integration of GIS and KDD in transport demand modelling. *Cartography and Geographic Information Science*, 31(1):19–28.

Ganskopp, D. (2001). Manipulating cattle distribution with salt and water in large arid-land pastures: a GPS/GIS assessment. *Applied Animal Behaviour Science*, 73(4):251–262.

Giraldeau, L. A.-A. and Caraco, T. (2000). *Social Foraging Theory*. Princeton University Press, Princeton, NJ.

Gold, C., Chau, M., Dziezko, M., and Goralski, R. (2004). 3d geographic visualization: The marine GIS. In Fisher, P. F., editor, *Developments in Spatial Data Handling*, pages 17–28. Springer, Berlin Heidelberg, DE.

Golledge, R. G. (2002). The nature of geographic knowledge. *Annals of the Association of American Geographers*, 92(1):1–14.

Goodchild, M. F. (1992). Geographical information science. *International Journal of Geographical Information Systems*, 6(1):31–45.

Goodchild, M. F. (2000). GIS and transportation: Status and challenges. *GeoInformatica*, 4(2):127–139.

Goodchild, M. F. (2004a). GIScience, geography, form, and process. *Annals of the Association of American Geographers*, 94(4):709–714.

Goodchild, M. F. (2004b). The validity and usefulness of laws in geographic information science. *Annals of the Association of American Geographers*, 94(2):300–303.

Goodrich, M. T. and Tamassia, R. (1998). *Data Structures and Algorithms in Java*. John Wiley & Sons, New York, NY.

Grumbach, S., Koubarakis, M., Rigaux, P., Scholl, M., and Skiadopoulos, S. (2003). Spatio-temporal models and languages: An approach based on constraints. In Koubarakis, M., Sellis, T., Frank, A. U., Grumbach, S., Gueting, R. H., Jensen, C. S., Lorentzos, N., Manolopoulos, Y., Nardelli, E., Pernici, B., Schek, H.-J., Scholl, M., Theodoulidis, B., and Tryfona, N., editors, *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, volume 2520 of *Lecture Notes in Computer Science*, chapter 5, pages 177–201. Springer, Berlin Heidelberg, DE.

Grumbach, S., Rigaux, P., and Segoufin, L. (2001). Spatio-temporal data handling with constraints. *GeoInformatica*, 5(1):95–115.

Gueting, R. H., Boehlen, M. H., Erwig, M., Jensen, C. S., Lorentzos, N., Nardelli, E., Schneider, M., and Vazirgiannis, M. (2000). A foundation for representing and querying moving objects. *ACM Transactions on Database Systems (TODS)*, 2520(1):1–42.

Gueting, R. H., Boehlen, M. H., Erwig, M., Jensen, C. S., Lorentzos, N., Nardelli, E., Schneider, M., and Viqueira, J. R. R. (2003). Spatio-temporal models and languages: An approach based on data types. In Koubarakis, M., Sellis, T., Frank, A. U., Grumbach, S., Gueting, R. H., Jensen, C. S., Lorentzos, N., Manolopoulos, Y., Nardelli, E., Pernici, B., Schek, H.-J., Scholl, M., Theodoulidis, B., and Tryfona, N., editors, *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, volume 2520 of *Lecture Notes in Computer Science*, chapter 4, pages 117–176. Springer, Berlin Heidelberg, DE.

Guibas, L. J. (1998). Kinetic data structures: A state of the art report. In Agarwal, K., Kavraki, L., and Mason, M., editors, *Proceedings of the third workshop on the algorithmic foundations of robotics on Robotics : the algorithmic perspective*, pages 191–209, Natrick, MA. A. K. Peters, Ltd.

Guibas, L. J. (2004). Motion. In Goodman, J. and O'Rourke, J., editors, *Handbook of Discrete and Computational Geometry*, chapter 50, pages 1117–1134. Chapman and Hall, CRC Press LLC, Boca Raton, FL, 2nd edition.

Hägerstrand, T. (1970). What about people in regional science. *Papers of the Regional Science Association*, 24:7–21.

Haller, R., Filli, F., and Imfeld, S. (2001). Evaluation of GPS-technology for tracking mountain ungulates: VHS-transmitters or GPS-collars? In Sibbald, A. M. and Gordon, I. J., editors, *Tracking Animals with GPS*, pages 61–66, Aberdeen, UK. Macaulay Insitute.

Hemelrijk, C. K. (2000). Towards the integration of social dominance and spatial structure. *Animal Behaviour*, 59(5):1035–1048.

Hermann, M. and Leuthold, H. (2001). Weltanschauungen und ihre soziale Basis im Spiegel der Volksabstimmungen. *Swiss Political Science Review*, 7(4):39–63.

Hermann, M. and Leuthold, H. (2003). *Atlas der politischen Landschaften. Ein weltanschauliches Portrait der Schweiz*. vdf-Verlag, Zurich, CH.

Hill, N. and Häder, D.-P. (1997). A biased random walk model for the trajectories of swimming micro-organisms. *Journal of theoretical Biology*, 186(4):503–526.

Hoare, D. J., Ruxton, G. D., Godin, J.-G. J., and Krause, J. (2000). The social organization of free-ranging fish shoals. *Oikos*, 89(3):546–554.

Hornsby, K. and Egenhofer, M. J. (2000). Identity-based change: a foundation for spatio-temporal knowledge representation. *International Journal of Geographical Information Science*, 14(3):207–224.

Hornsby, K. and Egenhofer, M. J. (2002). Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1-2):177–194.

Huisman, O. and Forer, P. (1998). Computational agents and urban life spaces: a preliminary realisation of the time-geography of student lifestyles. In *Proceedings of the 3rd International Conference of GeoComputation*. "GeoComputation CD-ROM".

Hulbert, I. A. R. (2001). Gps and its use in animal telemetry: The next five years. In Sibbald, A. M. and Gordon, I. J., editors, *Tracking Animals with GPS*, pages 51–60, Aberdeen, UK. Macaulay Insitute.

Imfeld, S. (2000). *Time, Points and Space: Towards a better analysis of wildlife data in GIS*. PhD thesis, University of Zurich.

Iwase, S. and Saito, H. (2002). Tracking soccer player using multiple views. In *IAPR Workshop on Machine Vision Applications*, pages 102–105, Tokyo, JP. MVA Proceedings.

Iwase, S. and Saito, H. (2003). Tracking soccer players based on homography among multiple views. In Ebrahimi, T. and Sikora, T., editors, *Visual Communications and Image Processing 2003*, volume 5150, pages 283–292. Proceedings of SPIE.

Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 22(1):4–37.

Jones, P. W. and Smith, P. (2001). *Stochastic Processes: An introduction*. Arnold Publishers, London, UK.

Jonsen, I. D., Myers, R. A., and Flemming, J. M. (2003). Meta-analysis of animal movement using state-space models. *Ecology*, 85(11):3055–3063.

Kääb, A. (2002). Monitoring high-mountain terrain deformation from repeated air- and spaceborne optical data: examples using digital aerial imagery and ASTER data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 57(1-2).

Kanellakis, P. C., Kuper, G. M., and Revesz, P. Z. (1995). Constraint query language. *Journal of Computer and System Sciences*, 51(1):26–52.

Knaden, M. and Wehner, R. (2003). Nest defense and conspecific enemy recognition in the desert ant *Cataglyphis fortis*. *Journal of Insect Behaviour*, 16(5):717–730.

Knaden, M. and Wehner, R. (2004). Path integration in desert ants controls aggressiveness. *Science*, 305(5680):60.

Knuth, D. E., Morris, J. H., and Pratt, V. R. (1977). Fast pattern matching in strings. *SIAM Journal on Computing*, 6(1):323–350.

Koubarakis, M. and Sellis, T. (2003). Introduction. In Koubarakis, M., Sellis, T., Frank, A. U., Grumbach, S., Gueting, R. H., Jensen, C. S., Lorentzos, N., Manolopoulos, Y., Nardelli, E., Pernici, B., Schek, H.-J., Scholl, M., Theodoulidis, B., and Tryfona, N., editors, *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, volume 2520 of *Lecture Notes in Computer Science*, chapter 1, pages 1–8. Springer, Berlin Heidelberg, DE.

Kraak, M.-J. and Koussoulakou, A. (2004). A visualization environment for the space-time-cube. In Fisher, P. F., editor, *Developments in Spatial Data Handling*, Proceedings ot the 11th International Symposium on Spatial Data Handling, pages 189–200. Springer, Berlin Heidelberg, DE.

Kraak, M.-J. and MacEachren, A. M. (1994). Visualisation of the temporal component of spatial data. In Waugh, T. C. and Healey, R. G., editors, *Advances in GIS research*, volume 1 of *Proceedings of the 6th International Symposium on Spatial Data Handling*, pages 391–405. Department of Geography, University of Edinburgh, Edinburgh, UK.

Krause, J. and Ruxton, G. D. (2002). *Living in Groups*. Oxford Series in Ecology and Evolution. Oxford University Press, New York, NY.

Kwan, M.-P. (2000). Interactive geovisualization of activity-travel patterns using three dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C*, 8(1-6):185–203.

Lachlan, R. F., Crooks, L., and Laland, K. N. (1998). Who follows whom? Shoaling preferences and social learning of foraging information in guppies. *Animal Behaviour*, 56(1):181–190.

Laube, P. and Imfeld, S. (2002). Analyzing relative motion within groups of trackable moving point objects. In Egenhofer, M. J. and Mark, D. M., editors, *Geographic Information Science*, volume 2478 of *Lecture Notes in Computer Science*, pages 132–144. Springer, Berlin Heidelberg, DE.

Laube, P., Imfeld, S., and Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668.

Laube, P. and Purves, R. S. (submitted 2005). Evaluating motion pattern detection techniques in spatio-temporal data. *Computers, Environment and Urban Systems*. Submitted April 2005.

Laube, P., van Kreveld, M., and Imfeld, S. (2004). Finding REMO - detecting relative motion patterns in geospatial lifelines. In Fisher, P. F., editor, *Developments in Spatial Data Handling*, Proceedings of the 11th International Symposium on Spatial Data Handling, pages 201–214. Springer, Berlin Heidelberg, DE.

Lee, H. and Kemp, Z. (1998). Supporting complex spatiotemporal analysis in GIS. In Carver, S., editor, *Innovations in GIS 5*, chapter 14, pages 151–161. Taylor & Francis, London, UK.

MacEachren, A. M. (1994). *Visualization in modern cartography.* Pergamon, Elsevier Science, Oxford, UK.

MacEachren, A. M. and Monmonier, M. (1992). Introduction to special issue on geographic visualization. *Cartography and Geographic Information Systems*, 19(4):197–200.

MacEachren, A. M., Wachowicz, M., Edsall, R., and Haug, D. (1999). Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science*, 13(4):311–334.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statisitics*. Wiley series in probability and statistics. John Wiley & Sons, New York, NY.

Mark, D. M. (1998). Geospatial lifelines. In *Integrating Spatial and Temporal Databases*, volume 98471. Dagstuhl Seminars.

Mark, D. M. (2003). Geographic information science: Defining the field. In Duckham, M., Goodchild, M. F., and Worboys, M. F., editors, *Foundations of Geographic Information Science*, chapter 1, pages 3–18. Taylor & Francis, London, UK.

Massey, D. (1999). Space-time, 'science' and the relationship between physical geography and human geography. *Transactions of the Institute of British Geographers*, 24(3):261–276.

Matthiopoulos, J., McConnell, B., Duck, C., and Fedak, M. (2004). Using satellite telemetry and aerial counts to estimate space use by grey seals around the British Isles. *Journal of Applied Ecology*, 41(3):476–491.

May, J. and Thrift, N. (2001). *TimeSpace: Geographies of Temporality*. Routledge, London, UK, and New York, NY.

McCormick, B., DeFanti, T., and Brown, M. (1987). Visualization in scientific computing. *ACM SIGGRAPH Computer Graphics*, 21(6).

McGrady, M. J., Ueta, M., Potapov, E. R., Utekhina, I., Masterov, V., Ladyguine, A., Zykov, V., Cibor, J., Fuller, M., and Seegar, W. S. (2003). Movements by juvenile and immature Steller's Sea Eagles *Maliaeetus pelagicus* tracked by satellite. *Ibis, Journal of the British Ornitological Union*, 145(2):318–328.

Miller, H. J. (1991). Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Systems*, 5(3):287–301.

Miller, H. J. (2003). What about people in geographic information science? *Computers, Environment and Urban Systems*, 27(5):447–453.

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2):284–289.

Miller, H. J. and Han, J. (2001). Geographic data mining and knowledge discovery: An overview. In Miller, H. J. and Han, J., editors, *Geographic data mining and knowledge discovery*, pages 3–32. Taylor & Francis, London, UK.

Miller, H. J. and Wentz, E. A. (2003). Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers*, 93(3):574–594.

Miller, H. J. and Wu, Y.-H. (2000). GIS software for measuring space-time accessibility in transportation planning and analysis. *GeoInformatica*, 4(2):141–159.

Morales, J. M. (2002). Behavior at habitat boundaries can produce leptokuritc movement distributions. *The American Naturalist*, 160(4):531–538.

Morales, J. M. and Ellner, S. P. (2002). Scaling up animal movements in heterogeneous landscapes: The importance of behavior. *Ecology*, 83(8):2240–2247.

Moreira, J., Ribeiro, C., and Saglio, J.-M. (1999). Representing and manipulation of moving points: An extented data model for location estimation. *Cartography and Geographic Information Science*, 26(2):109–123.

Mountain, D. and Raper, J. (2000). Designing geolocation services for next generation mobile phone systems. Technical report, GIS 2000 in association with the AGI Conference.

Mountain, D. and Raper, J. (2001a). Modelling human spatio-temporal behaviour: A challenge for location-based services. In *Proceedings of the 6th International Conference of GeoComputation*. "GeoComputation CD-ROM".

Mountain, D. and Raper, J. (2001b). Spatio-temporal representations of individual human movement for personalising location based services. In Kidner, D. B. and Higgs, G., editors, *Proceedings of the GIS research UK, 9th Annual Conference, GISRUK 2001*, pages 579–582. University of Glamorgan, UK.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Geo Books, Norwich, UK.

O'Sullivan, D. and Unwin, D. J. (2003). *Geographic Information Analysis*. John Wiley & Sons, Hoboken, NJ.

Padmanabhan, B. (2004). The interestingness paradox in pattern discovery. *Journal of Applied Statistics*, 31(8):1019–1035.

Pang, M. Y. C. and Shi, W. (2002). Development of a process-based model for dynamic interaction in spatio-temporal GIS. *GeoInformatica*, 6(4):323–344.

Pasquale, A. D., Forlizzi, L., Jensen, C. S., Manolopoulos, Y., Nardelli, E., Pfoser, D., Proietti, G., Saltenis, S., Theodoridis, Y., Tzouramanis, T., and Vassilakopoulos, M. (2003). Access methods and query processing techniques. In Koubarakis, M., Sellis, T., Frank, A. U., Grumbach, S., Gueting, R. H., Jensen, C. S., Lorentzos, N., Manolopoulos, Y., Nardelli, E., Pernici, B., Schek, H.-J., Scholl, M., Theodoulidis, B., and Tryfona, N., editors, *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, volume 2520 of *Lecture Notes in Computer Science*, chapter 6, pages 203–261. Springer, Berlin Heidelberg, DE.

Pearson, H., Clarke, T., Abbott, A., Knight, J., and Cyranoski, D. (2003). SARS - what have we learned. *Nature*, 424:121–126.

Peterson, M. P. (1995). *Interactive and Animated Cartography*. Prentice Hall, Englewood Cliffs, NJ.

Peterson, R. O., Jacobs, A. K., Drummer, T. D., Mech, L. D., and Smith, D. W. (2002). Leadership behavior in relation to dominande and reproductive status in gray wolves, *Canis lupus*. *Cana-*

dian Journal of Zoology, 80(8):1405–1412.

Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 83(3):441–461.

Peuquet, D. J. (2001). Making space for time: Issues in space-time data representation. *GeoInformatica*, 5(1):11–32.

Peuquet, D. J. (2002). *Representation of Space and Time*. The Guilford Press, London, UK.

Peuquet, D. J. and Duan, N. (1995). An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographic data. *International Journal of Geographical Information Systems*, 9(1):7–24.

Peuquet, D. J. and Marble, D. F. (1990). *Introductory readings in geographic information systems*. Taylor & Francis, London, UK.

Pfoser, D. and Jensen, C. S. (1999). Capturing the uncertainty of moving-object representations. In Gueting, R. H., Papadias, D., and Lochowsky, F., editors, *Advances in Spatial Databases, 6th International Symposium, SSD'99, Hong Kong, China, July 1999. LNCS 1651*, pages 111–131. Springer, Berlin Heidelberg, DE.

Pfoser, D. and Theodoridis, Y. (2003). Generating semantics-based trajectories of moving objects. *Computers, Environment and Urban Systems*, 27(3):243–263.

Ramos-Fernández, G., Mateos, J. L., Miramontes, O., Cocho, G., Larralde, H., and Ayala-Orozco, B. (2004). Lévy walk patterns in the foraging movement of spider monkeys (*Ateles geoffroyi*). *Behavioral Ecology and Sociobiology*, 55(3):223–230.

Rana, S. and Dykes, J. (2003). A framework for augmenting the visualization of dynamic raster surfaces. *Information Visalization*, 2(2):126–139.

Rands, S. A., Cowlishaw, G., Pettifor, R. A., Rowcliffe, J. M., and Johnstone, R. A. (2003). Spontaneous emergence of leaders and followers in foraging pairs. *Nature*, 423(6938):432–434.

Raper, J. (2000). *Multidimensional Geographic Information Science*. Taylor & Francis, London, UK.

Raper, J. (2002). The dimensions of GIScience. Keynote at 2nd International Conference of Geographic Information Science, GIScience 2002, Boulder, CO, USA, September 25-28.

Raper, J. and Livingstone, D. (1995). Development of a geomorphological spatial model using object-oriented design. *International Journal of Geographical Information Systems*, 9(4):359–383.

Raptopoulou, K., Papadopoulos, A. N., and Manolopoulos, Y. (2003). Fast nearest-neighbor query processing in moving-point databases. *GeoInformatica*, 7(2):113–137.

Robinson, A. H., Morrison, J. L., Muercke, P. C., Kimerling, A. J., and Guptil, S. C. (1995). *Elements of Cartography.* John Wiley & Sons, New York, NY, 6th edition.

Sadahiro, Y. (2002). A graphical method for exploring spatiotemporal point distributions. *Cartography and Geographic Information Science*, 29(2):67–84.

Schmitt, F. G. and Seuront, L. (2001). Multifractal random walk in copepod behaviour. *Physica A*, 301(1-4):375–396.

Sears, F. W., Zemansky, M. W., and Young, H. D. (1987). *University Physics.* Addison-Wesley, Reading, MA, 7th edition.

Seeley, T. D. and Visscher, P. K. (2004). Group decision making in nest-site selection by honey bees. *Apidologie*, 35(2):101–116.

Sibbald, A. M., Hooper, R. J., Gordon, I. J., and Cumming, S. (2001). Using GPS to study the effect of human disturbance on the behaviour of the red deer stags on a highland estate in Scotland. In Sibbald, A. M. and Gordon, I. J., editors, *Tracking Animals with GPS*, pages 39–43, Aberdeen, UK. Macaulay Insitute.

Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974.

Sistla, A. P., Wolfson, O., Chamberlain, S., and Dao, S. (1997). Modeling and querying moving objects. In *Proceedings of the 13th International Conference on Data Engineering (ICDE13).*

Sistla, A. P., Wolfson, O., Chamberlain, S., and Dao, S. (1998). Querying the uncertain position of moving objects. In Etzion, O., Jajodia, S., and Sripada, S., editors, *Temporal Databases - Research and Practice*, volume 1399 of *Lecture Notes in Computer Science*, pages 310–337. Springer, Berlin Heidelberg, DE.

Smyth, C. S. (2001). Mining mobile trajectories. In Miller, H. J. and Han, J., editors, *Geographic data mining and knowledge discovery*, pages 337–361. Taylor & Francis, London, UK.

Summers, C. (2003). Mobile phones – the new fingerprints.

Suter, M. (2001). *Richtig leben mit Geri Weibel.* Diogenes, Zurich, CH.

Theodoridis, Y. and Papadias, D. (1995). Range queries involving spatial relations: A performance analysis. In Frank, A. U. and Kuhn, W., editors, *Proceedings of the 2nd International Conference of Spatial Information Theory (COSIT'95)*, volume 988 of *Lecture Notes in Computer Science*, pages 537–551. Springer, Berlin Heidelberg, DE.

Thériault, M., Claramunt, C., and Villeneuve, P. Y. (1999). A spatio-temporal taxonomy for the representation of spatial set behaviours. In Böhlen, M. H., Jensen, C. S., and Scholl, M. O., editors, *Spatio-Temporal Database Management*, volume 1678 of *Lecture Notes in Computer Science*, pages 1–18. Springer, Berlin Heidelberg, DE.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240.

Trajcevski, G., Wolfson, O., Hinrichs, K., and Chamberlain, S. (2004). Managing uncertainty in moving objects databases. *ACM Transactions on Database Systems (TODS)*, 29(3):463–507.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Turchin, P. (1998). *Quantitative Analysis of Movement: Measuring and Modelling Population Redistribution in Animals and Plants*. Sinauer Publishers, Sunderland, MA.

Turner, M. G., Wu, Y., Wallace, L. L., Romme, W. H., and Brenkert, A. (1994). Simulating winter interactions among ungulates, vegetation, and fire in northern Yellowstone Park. *Ecological Applications*, 4(4):472–496.

Tuyttens, F. A. M., Macdonald, D. W., and Roddam, A. W. (2002). Effects of radio-collars on European badgers (*meles meles*). *Journal of Zoology*, 257(1):37–42.

University Consortium for Geographic Information Science (1996). Research priorities of geographic information science. *Cartography and Geographic Information Systems*, 23(3):115–127.

von Hühnerbein, K., Hamann, H., Ritter, E., and Wiltschko, W. (2000). A GPS-based system for recroding the flight paths of birds. *Naturwissenschaften*, 87(6):278–279.

Wachowicz, M. and Healey, R. G. (1994). Towards temporality in GIS. In *Innovations in GIS 1*, chapter 8, pages 105–115. Taylor & Francis, London, UK.

Watanabe, S. (1985). *Pattern Recognition: Human and Mechanical*. John Wiley & Sons, New York, NY.

Wentz, E. A., Campell, A. F., and Houston, R. (2003). A comparison of two methods to create tracks of moving objects: linear weighted distance and constrained random walk. *International Journal of Geographical Information Science*, 17(7):623–645.

Westervelt, J. D. and Hopkins, L. D. (1999). Modeling mobile individuals in dynamic landscapes. *International Journal of Geographical*

Information Science, 13(3):191–208.

Wolfson, O., Jiang, L., Sistla, A. P., Chamberlain, S., Rishe, N., and Deng, M. (1998a). Databases for tracking mobile units in real time. In Beeri, C. and Buneman, P., editors, *Proceedings of the Seventh International Conference on Database Theory (ICDT), Jerusalem, Israel, Jan. 10-12, 1999*, number 1540 in Lecture Notes in Computer Science, pages 169–186. Springer, Berlin Heidelberg, DE.

Wolfson, O., Sistla, A. P., Chamberlain, S., and Yesha, Y. (1999). Updating and querying databases that track mobile units. *Distributed and Parallel Databases*, 7(3):257–287.

Wolfson, O., Xu, B., Chamberlain, S., and Jiang, L. (1998b). Moving objects databases: Issues and solutions. In Rafanelli, M. and Jarke, M., editors, *10th International Conference on Scientific and Statistical Database Management, Proceedings, Capri, Italy, July 1-3, 1998*, pages 111–131. IEEE Computer Society.

Wood, J. (2002). *Java Programming for Spatial Sciences*. Taylor & Francis, London, UK.

Worboys, M. F. (1998). A generic model for spatio-bitemporal geographic information. In Egenhofer, M. J. and Colledge, R. G., editors, *Spatial and Temporal Reasoning in Geographic Information Systems*, chapter 2, pages 25–39. Oxford University Press, Oxford, UK.

Part II

Publications

Credits and Permissions

Attached are the four publications building the scientific core of this thesis. The first three papers are attached in the originally published format, and full credits and reproduction permissions are provided below. Since the fourth paper is only submitted, it has a preprint format.

Laube, P. and Imfeld, S. (2002). Analyzing relative motion within groups of trackable moving point objects. In Egenhofer, M. J. and Mark, D. M., editors, *Geographic Information Science*, volume 2478 of *Lecture Notes in Computer Science*, pages 132–144. Springer, Berlin-Heidelberg, DE. **Reproduced with kind permission of Springer Science and Business Media, 22 June 2005.**

Laube, P., Imfeld, S., and Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668. **Reproduced with kind permission of Taylor & Francis, 15 June 2005.**

Laube, P., Van Kreveld, M. and Imfeld, S. (2004). Finding REMO - detecting relative motion patterns in geospatial lifelines. In Fisher, P. F., editor, *Developments in Spatial Data Handling*, Proceedings of the 11th International Symposium on Spatial Data Handling, pages 201–215. Springer, Berlin-Heidelberg, DE. **Reproduced with kind permission of Springer Science and Business Media, 22 June 2005.**

Laube, P. and Purves, R.S (submitted 2005). Evaluating motion pattern techniques in spatio-temporal data. *Computers, Environment and Urban Systems*, Submitted April 2005.

Analyzing Relative Motion within Groups of Trackable Moving Point Objects

Patrick Laube and Stephan Imfeld

Geographic Information Systems Division,
Department of Geography, University of Zurich
Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
{plaube,imfeld}@geo.unizh.ch

Abstract. The overall goal of the ongoing project is to develop methods for spatio-temporal analysis of relative motion within groups of moving point objects, such as GPS-tracked animals. Whereas recent efforts of dealing with dynamic phenomena within the GIScience community mainly concentrated on modeling and representation, this research project concentrates on the analytic task. The analysis is performed on a process level and does not use the traditional cartographic approach of comparing snapshots. The analysis concept called REMO (RElative MOtion) is based on the comparison of motion parameters of objects over time. Therefore the observation data is transformed into a 2.5-dimensional analysis matrix, featuring a time axis, an object axis and motion parameters. This matrix reveals basic searchable relative movement patterns. The current approach handles points in a pure featureless space. Case study data of GPS-observed animals and political entities in an ideological space are used for illustration purposes.

1 Introduction

The only constant in the contemporary world is change. Nothing is ever stable. Sciences investigate change to understand the underlying processes, to find universal rules. The collection of data, the descriptive part of scientific activities, is necessary to have the foundations on which deductive work is building up [2]. According to Frank change comes in two forms: change of the objects of interest (*life*) and change in the position or geometric form of these objects (*motion*). Motion, the topic of this paper, is a spatio-temporal phenomenon.

Space is crucial to investigate motion, because a change of position can only be noticed in relation to a reference, mostly a spatial coordinate system. *Time* is crucial too as it is intrinsically linked with causation. Causation implies precedence, lack of precedence rules out causation [2].

Geographic information systems (GISs) provide a wide range of analysis tools for spatial science. Their potential to deal with spatio-temporal phenomena such as change is not yet very elaborate. In the late eighties the topic of time entered the field of GIScience. Langran identified the need to describe spatial change over time and examined the design of temporal GIS [8]. Hornsby and Egenhofer [5] present an approach to represent and model geographic entities in time. Peuquet [11] gives a detailed overview on today's issues in space–time data representation and tries to ex-

M.J. Egenhofer and D.M. Mark (Eds.): GIScience 2002, LNCS 2478, pp. 132–144, 2002.

plain, why the representation of both space and time in digital databases is still problematic.

Examples for analysis of change on a process level are sparse in literature. Openshaw *et al.* [10] and Imfeld [7] give two examples. Openshaw *et al.* present an approach for spatio-temporal data mining. The Space-Time-Attribute Analysis Machine (STAM) and Space-Time-Attribute Creature (STAC) try to find clusters and other patterns in space and/or time in long-term illness census data without prior knowledge. Imfeld presents methods termed the time plot family to analyze mobile objects in their environment. In his approach the movements of up to two individuals are investigated in an analysis space featuring two temporal axes.

The classical situation for motion is humans or animals moving in space. These phenomena can easily be modeled and handled as a series of observations of moving points, represented as tuples of t, x, y and z coordinates. Due to substantial advances in tracking technologies, such as GPS and mobile device technologies, an increase in spatio-temporal data about moving objects can be expected. In the field of animal telemetry GPS technology provides observation data of previously unseen qualities and quantities [6].

In the late seventies, Bertin used the procedure of seriation to find behavioral patterns in animal observation data [1, 9]. Seriation is an experimental tool for ordering and classifying two-dimensional tables relating to sets of elements. The rows and columns of a matrix are permutated such that, starting from any specific element, the other elements most similar to it are closest in the sequence. In one example he found behavioral response of woodlouses to a light source by ordering motion parameters using paper file cards.

Observations of moving point objects are not only seen in wildlife sciences, but also in other domains like social sciences, geomarketing, transport GIS (TGIS) or even the political sciences. Haggett's famous people on the beach illustrate such a social phenomenon [3]. Also, in social sciences political entities such as communities can be plotted over time in abstract ideological spaces, for instance between the extremes left, right, conservative and progressive [4]. A telecommunication company might be interested in the spatio-temporal behavior of cellular phone users for public relations or network expansion planning. Radio-tracked taxicabs in a city are an example for TGIS [12]. All those phenomena can be reduced to the basic phenomena of moving points and then might be analyzed by similar spatio-temporal analysis methods.

The overall goal of the research presented in this paper is to find, quantify and visualize user-defined motion patterns in groups of moving point objects. The detailed aims are:

- To develop a flexible analysis concept for the integrated analysis of motion parameters of groups of moving point objects.
- To identify, characterize and categorize the basic types of relative motion within groups of moving point objects.
- To identify (sub–)groups according to equal or similar movements.
- To develop algorithms to recognize patterns in the data. Finding patterns over time means identifying (a) the concerned individuals and (b) their location and extent on the time axis.

2 Analysis Concept

The basic idea of the analysis concept presented here is to compare the motion parameters of different point objects over space and over time. Thus, the fundamental concept underlying the analysis concept can be called "relative motion," giving the analysis concept its name: REMO (*RElative MOtion*).

The concept is based on the combination of the two ideas of arranging the tracking data tuples and detecting discrete motion patterns.

- For large spatio–temporal data sets systematic arranging is a key concept to gain insight. The first key approach is to transform the tracking data into arrangement or a configuration that reveals motion patterns (section 2.1).
- Basing on this arrangement, a second important approach is to search discrete instances of change rather than trying to perceive the entire motion processes. Analyzing incidents of delimited change is easier than analyzing the processes themselves. Every complex motion behavior can be fractionalized to discrete behavioral patterns, such as "sudden change in motion direction" or "many objects moving with equal speed" (section 2.2)

2.1 The REMO Analysis Matrix

Let *N* be a set of individual point objects moving around (Fig. 1a). The four tracks could represent the observations of four caribou cows carrying GPS-collars (O_1 to O_4). The motion of every object is recorded as a path of exact coordinate tuples with the structure t, X_t, Y_t, Z_t. The recordings of the positions of the point objects must be performed synchronously (t_1 to t_5). Thus, at every time step a set of motion parameters can be derived from the original tuples, taking into account at least two consecutive observations. The three basic motion parameters are *motion azimuth*, *speed* and *change of speed (∂-speed)*.

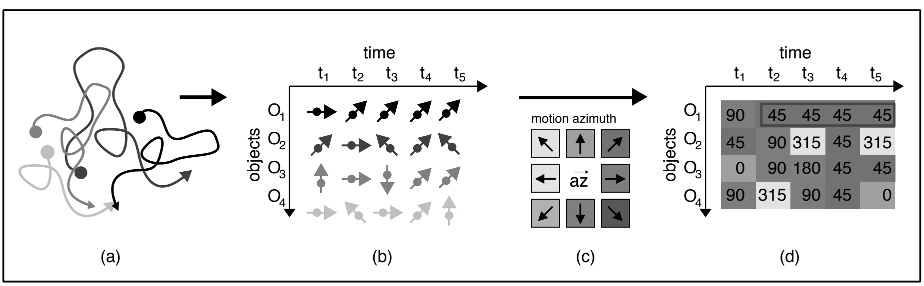


Fig. 1. The construction of the REMO analysis matrix

The motion parameter values are arranged in a matrix (e.g., motion-azimuth in Fig. 1b). Every row of the matrix represents the sequence of an object's motion parameter values in regular intervals over time. The rows with the object's motion parameter values are arranged so that coinciding observations are vertically aligned. Thus, the columns of the matrix represent time steps, specifying the object's temporally coinciding motion parameters. This matrix can be considered as a 2.5-dimensional analysis space: The horizontal axis is defined as the temporal axis (*time t)*, the vertical axis is defined as the *object axis* and the values represent the *motion parameter values*. While the temporal axis is ordered, the object axis includes no explicit order among the objects. The REMO analysis concept is constructed such that no a priori ordering among the objects is needed.

The REMO analysis concept compares the motion parameters of different point objects at different times. Therefore the motion parameters are grouped into discrete classes. For the azimuth eight classes have been chosen (N, NE, E, SE, S, SW, W, and NW, Fig. 1c). Also speed and ∂-speed have been reclassified: Eight classes of speed reduced to the unit interval 0 (slowest) to 1 (fastest) and eight classes of ∂-speed reduced to the unit interval—1 (maximum slowing-down) to +1 (maximum speeding-up).

In the REMO analysis concept space is continuous. In the strict sense the temporal dimension is discrete. However, since the temporal steps between the observations are kept short (in relation to the whole monitoring time frame) time shall also be considered as continuous. The temporal coverage of the motion is assumed to be complete. Missing observations might be interpolated.

The REMO matrix with the numerical description of the object's motion is the basis of all further analyses. A straightforward approach is to visualize the sequences of classified motion parameters in a halftone-coded matrix (Fig. 1d).

2.2 Basic Motion Patterns

The aim of arranging the motion parameters in a matrix is finding interrelations in the motion of a group of point objects. It is assumed that interrelations among the moving objects are manifested as patterns in the REMO matrix. Sequences and incidents that are somehow clustered on the time-axis and across the objects build a pattern in the REMO matrix. A pattern found in the REMO matrix could stand for a causal relation among the objects movements and initiates further investigation in the semantics of the observations.

What is meant by the term *pattern* in the REMO analysis concept? A pattern is a search template (e.g., a sequence of four times a motion azimuth of 45° found in the top row in Fig. 1d). A pattern is a defined set of motion parameter values with an extent in time and/or across the objects. Thus, patterns can span over the following two *dimensions* or their combination (Fig. 2).

- *(C) Patterns over time* (several times, one object (t:1); parallel to t-axis): Comparing the motion parameter values of one object at several times. Patterns may be cyclical or intermittent changes or trends in motion parameter values. Example: An object moves for four temporal intervals in the same direction.
- *(B) Patterns across objects* (same time, several objects (1:n); perpendicular to t-axis): Comparing the motion parameter values of the considered objects at a certain time. A pattern across objects is found, when a set of objects perform the same motion at the same time. Example: Five objects are moving in the same direction at time t.

- *(C) Combined patterns over time and across objects* (several times, several objects (t:n), both dimensions): The consideration of interrelation not only in one, but in both dimensions of the matrix discovers complex interactions between the motion of many objects at several time steps. Example: Object O_x anticipates the motion direction of four other objects.

Patterns within one dimension are called *simple* (A, B), those over time and across objects are called *complex* (C). A pattern has a temporal extent called the *pattern duration* with a start–time, duration, and an end–time. At the object axis the range can vary from only one to all objects concerned. The number of involved objects is called the *pattern width*.

Patterns can be discontinuous. *Gaps* in the patterns are of thematic interest and thus explicitly allowed. In the temporal dimension a gap in a pattern is needed to describe *time lags*, for instance between the cause and the effect of a phenomenon. Gaps in the object dimension arise out of the unordered construction of the matrix.

Two objects performing for instance the same abrupt change in motion direction may in real world space be positioned right next to each other or separated by a long distance. They are performing the same motion but in two different spatial contexts. Thus, the motion pattern is not the same and must be distinguished. The REMO analysis can be performed in two ways:

- *Motion Patterns without Neighborhood Information*. All objects are treated equally, no matter where they are in space, no matter whether they are close neighbors or have any other spatial interaction. Thus, in all these approaches the spatial dimension is reduced to the points' current motion parameters. The ordering dimension is time.
- *Motion Patterns using Neighborhood Information*. The absolute and relative positions of the moving objects are taken into account. This may happen as a preselection, based on proximity or other spatial interrelation.

Motion Patterns without Neighborhood Information. This section introduces a selection of the basic relative motion patterns in the REMO analysis concept (Fig. 2). The following analysis tasks are performed from a non-spatial perspective, all objects in the sample are taken into consideration, and the spatial relations among the objects are excluded. Thus, neighborhood does not matter.

Patterns over Time. Analysis over time searches for motion patterns parallel to the t-axis. These patterns are called SEQUENCE. To build a pattern over time, a minimum of two consecutive observations of one object is needed. Thus, the time frame lies between an interval of at least two consecutive observations and the extreme of the integrated analysis of all observations.

- The simplest and most obvious pattern is *constance*. The task is to find intervals with constant motion parameter values in an object's history (see grey box in Fig. 1d).
- A second sequence pattern is *turn*. A turn is a defined change in an object's motion parameter. An obvious turn is the alternation of the motion direction. Turns have (a) a parameter extent (value range) and (b) a temporal extent (duration).

Patterns across Objects. Analysis across objects compares the motion parameter values of a set of objects at certain moments in time. The task is finding INCIDENTS. In the REMO analysis concept an incident is a pattern in the motion parameter values of a set of moving objects that delimits this moment from the rest of the observation.

- *Concurrence* is the basic concept of patterns across objects, similar to constance in patterns over time. A concurrence pattern is found, when a set of objects (e.g., 60% of the whole sample) show a synchronous or at least similar motion parameter values at a certain time.
- The pattern *opposition* describes a bi- or multi-polar arrangement of motion parameter values. A typical case of opposition is the spatial splitting of a group of moving objects shown in a sudden appearance of two opposite motion directions (*bimodality*).
- The opposite of concurrence is *dispersion*. Dispersion can be an evident pattern in a group of moving point objects that is performing a non-uniform or random motion.

Complex Patterns over Time and across Objects. The motion patterns of the previous sections are simple and extend either along or perpendicular to the time axis. The patterns in this section combine motion patterns along and perpendicularly to the time axis of the REMO matrix, trying to find INTERACTIONS between the sequences of one object and incidents among the others.

- The most obvious complex analysis task is seeking for *trendsetters*. The basic idea is to find objects that anticipate a certain pattern of motion parameters that is afterwards reproduced by a set of the other objects. Thus, the complex pattern trendsetter is a combination of the simple patterns concurrence and constance.
- A pattern similar to the trendsetter can be called the *independent*: This pattern can be found if an individual moving point object goes his own way, ignoring the movement of the other objects of the sample.
- Another complex pattern is *propagation*. One objects starts to show a certain motion parameter value, and little by little other objects take part. With every time step more objects are involved.

The simple concept of concurrence can easily be extended along the time axis. The search template is simply extended to a temporal interval, in which a set of objects shows the same or at least similar motion parameter values. There are two basic patterns belonging to this complex pattern category.

- The first, *group turn* is a sudden change in the motion parameter values of a whole group of moving objects.
- The second is *group concurrence*. For this pattern a defined group of moving point objects shows a synchronous motion over a certain temporal interval, showing heterogeneous motion before and after.

Motion Patterns Using Neighborhood Information. Assuming that the objects of a group perform their motion in relation to the movement of other group members, the motions of the point objects must also be analyzed in their spatial context. The convergence of many caribou cows to the calving site, for example, can not be seen only in the parameters motion azimuth, speed and ∂-speed. To recognize this process also the motion of the objects in absolute space must be considered. The question is: Do objects that are showing a relative motion pattern have any spatial interrelation?

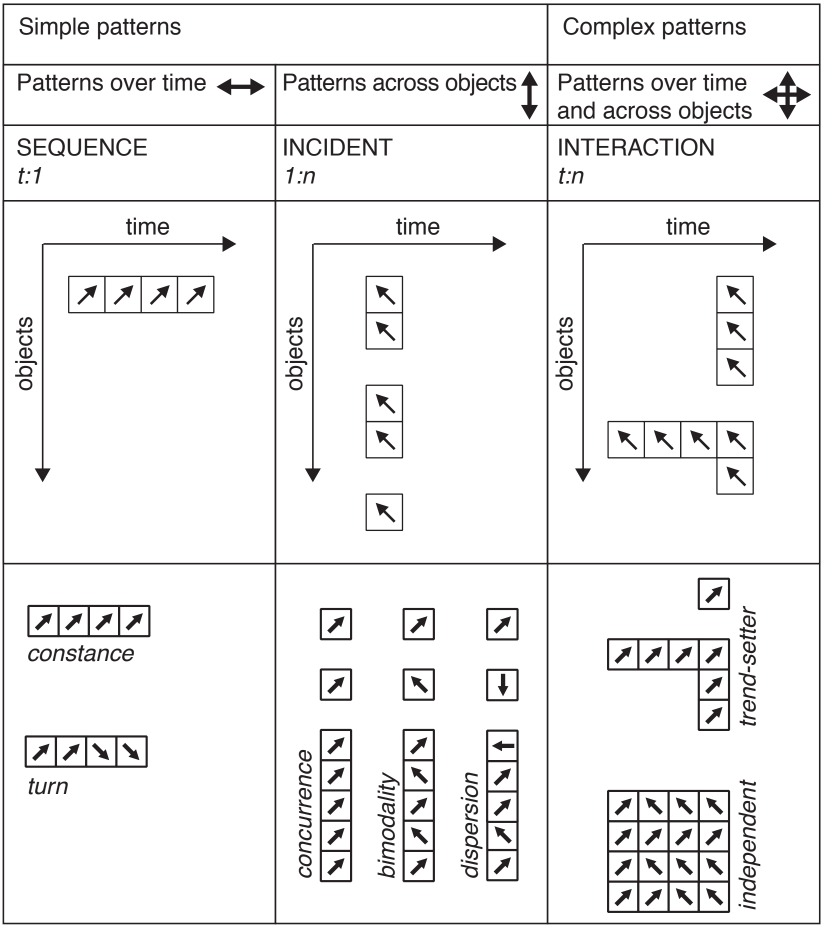


Fig. 2. Basic types of relative motion patterns

A variety of measures can describe the location of an individual relative to the others. They can be summarized as CENTRALITY/PERIPHERY MEASURES.

- The archetype of this kind of patterns can be called a *flock*. A flock is a point cluster showing concurrence (i.e., a conform motion). Measures need to be defined or adapted that indicate the presence of clusters in the REMO matrix. This opens a connection to the broad field of cluster analysis.
- The relation of *convergence* is found when a group of point objects is simultaneously heading for an identifiable point in space.
- The opposite relation *divergence* describes a group that disperses.

These patterns can only be detected by considering simultaneously the relative and absolute motions of objects in space and time.

The concept of the trendsetter excludes deliberately the relative position of the objects to each other in real space. In consideration of spatial interaction, the pattern trendsetter can be broken down further into the patterns *geese flock* and *guru*.

- In the case of *geese flock* the trendsetter shows a spatial proximity to its followers.
- The *guru* shows its trend-setting movement spatially apart from its followers.

3 Test Data Sets

The ultimate purpose of this research is the development of tools that allow quantitative analysis, not visualization that leads to qualitative analysis. Even though visualization is not the prime research objective, it is a powerful development approach. Thus, this conceptual paper illustrates the principles of the REMO analysis concept using visualizations of real data. The motion parameters azimuth, speed and ∂-speed of two test data sets are plotted in a black and white halftone-coded matrix (see online version for color pictures).

3.1 The Porcupine Caribou Project

The Porcupine Caribou Herd Satellite Collar Project is a cooperative project that uses satellite radio collars to document seasonal range use and migration patterns of a Porcupine Caribou herd in northern Yukon, Alaska and NWT. At the start of this project, 10 cow caribou were captured in October and November 1997 and equipped with GPS collars. Since then the caribou were tracked in approximately weekly intervals. (Details about the Porcupine Caribou Project including the original tracking data can be found at http://www.taiga.net/satellite/index.html).

Looking at the halftone–coded matrix an annual migration structure can be seen (Fig. 3). During the spring months March to May the N– and NE–heading azimuths dominate (light-grey). In the second half of the years rather the S– and E–Azimuths appear to prevail (dark-grey). While the winter and spring months show mainly constant slow movements, summer and fall see a more complex pattern of slow and fast movements, speeding-up and slowing-down, respectively.

Going into detail, the plot pinpoints some basic relative motion patterns. A phase of highly synchronous movements in June and July 1999 strikes as an obvious feature. This phase includes several *concurrence* patterns in the azimuth plot with pattern widths of up to 5 of 7 individuals. *Constance* patterns are pretty common in all plots, especially in the springtime. In June and July 1999 a *group turn* pattern from NW to SE is very obvious.

3.2 Swiss Political Districts Moving in an Abstract, Ideological Space

The frequently held popular referendums in Switzerland allow people to make detailed inferences about value–conflicts within the society. Hermann and Leuthold [4] developed an inductive approach to discover the basic ideological conflicts in Switzerland. Performing factor analysis on referendum data at the district level of all 158 federal referendums held between 1981 and 1999, they discovered a structure of mentality, which is composed of three dimensions: left vs. right, liberal vs. conservative and ecological vs. technocratic. The projections of this multidimensional ideological space, taken in pairs, provide a total of three two–dimensional maps of the political landscape of Switzerland.

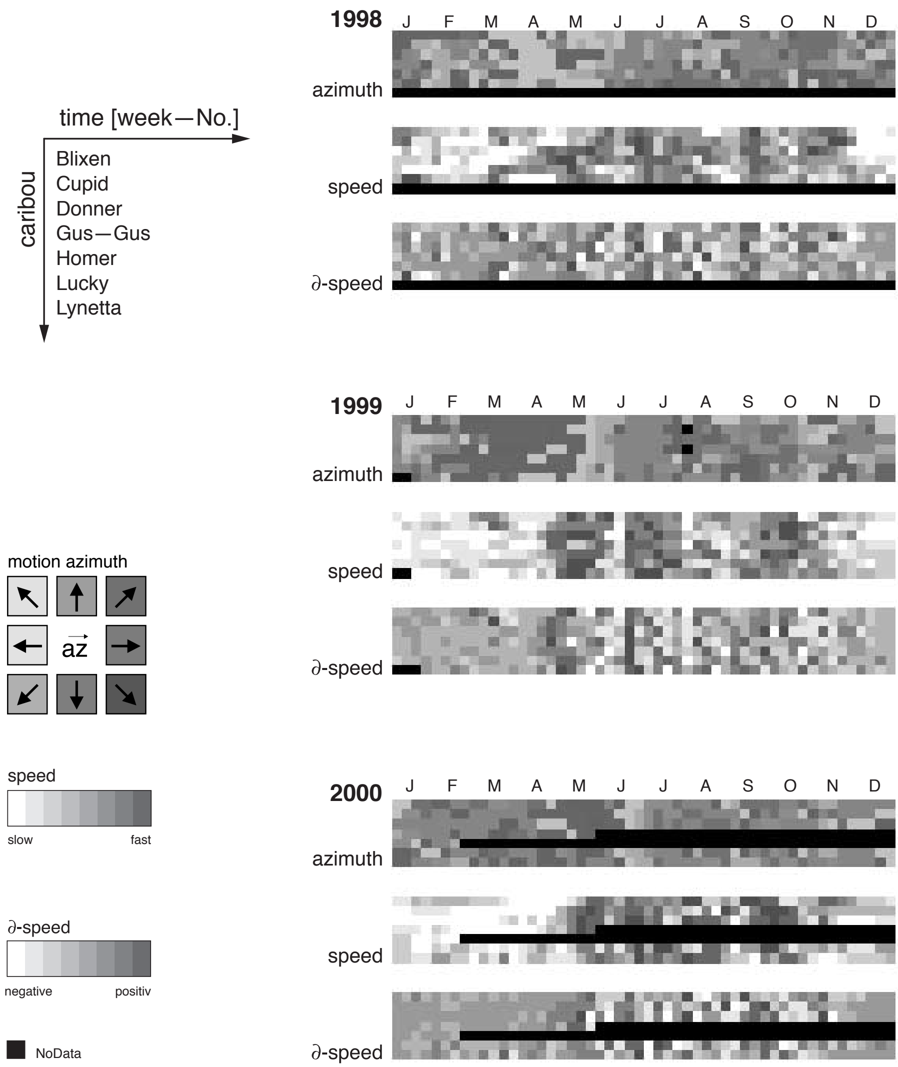


Fig. 3. REMO-Matrix of selected individuals of the Porcupine Caribou Project

In these two–dimensional ideological spaces the 185 districts can be localized in intervals of one year, from 1981 until 1999. Irrespective of their political and social meaning, the districts can be considered as moving points in a two–dimensional space. Thus, this is a data set that is suited for the REMO analysis concept.

(An animated view of the Swiss districts moving in the ideological space can be seen under the URL: http://www.geo.unizh.ch/gia/research/sotomo/gifs/filmCH.gif).

All districts of a canton are grouped together in the matrix (Fig. 4). Thus, proximity in the matrix corresponds to a certain institutional and cultural similarity.

The first overview allows an obvious distinction between the German speaking (topmost two fifths of districts) and the French and Italian speaking, Latin part (lowermost two fifth of districts) of Switzerland. Whereas the German speaking part in general moves to the right, the Latin part moves to the opposite left pole (azimuth). The French and German speaking, bilingual Canton of Fribourg (FR) endorses this general impression. Regardless of its separation from the other Latin Cantons by lower district numbers, it shows a Latin style color pattern.

By the means of the lexically ordered districts, the *concurrency* sequence of many cantons is apparent. Obvious examples are the cantons Ticino (TI), Vaud (VD), and Zurich (ZH) showing very comparable patterns through all districts and trough all three plots. Seeking for other basic REMO patterns *constance* strikes the most. In the azimuth plot many districts show persistence in their motion azimuth class for up to ten and more time steps. Also *turn* patterns are often seen. The canton of Zurich (ZH) shows some very obvious examples in the period 1985 to 1989. Most of its districts perform a directional turn from “left–technocratic” over “technocratic” to “right” in only four time steps (i.e., four years respectively). The initial starting point for the sequence may vary up to three years, but the overall azimuth change pattern remains the same. Another basic pattern can be seen in the districts of the Canton of Zurich. The district of Meilen (No. 7 from the top) anticipates the azimuth change of almost all other districts: Already in the early nineties it performs the directional change from “right” to “right–ecological.” The rest of the districts follow after 1992 one by one.

4 Discussion

The integrated investigation of the motion parameters motion azimuth, speed and ∂-speed provided new insights in the large and complex data sets. The possibility to be simultaneously aware of different motion parameters over any desired time frame allows a new way of investigating processes within groups of moving points.

The REMO analysis concept allows the identification of (sub-)groups showing equal or similar motion parameter values. Group identification was moderately possible within only a dozen individuals in the porcupine caribou data and explicitly possible in the “moving Swiss districts” data.

The two prototypical examples of the porcupine caribou data and the Swiss political districts data show that visualization is a powerful tool for the REMO analysis concept and thus must be implemented in an integrated analysis framework. But these examples also reveal the limits of analysis based exclusively on the halftone–coded matrix visualization. The larger the data sets are, the more difficult it is to interpret the visualization of the halftone–coded matrix. For larger data sets it is essential to develop precise measures and selective tools to identify events, processes and

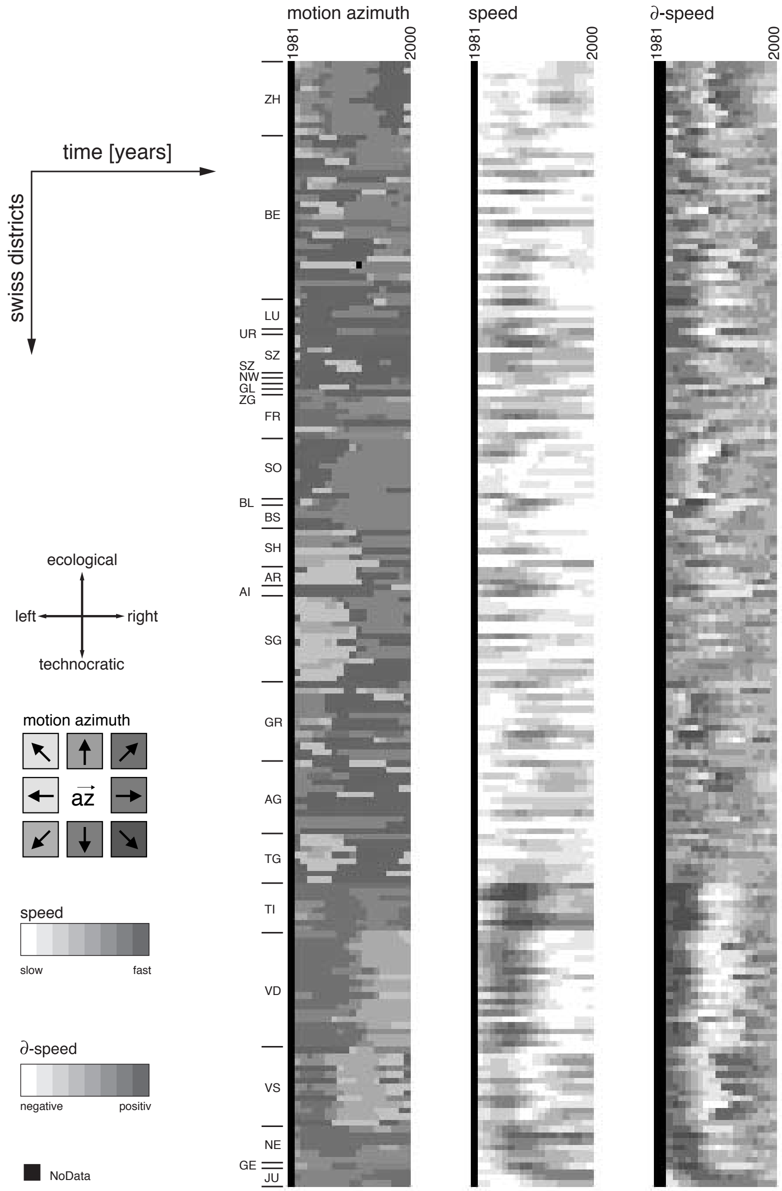


Fig. 4. REMO-Matrix of the Swiss political districts moving in an ideological space

relationships in a quantitative and reproducible manner. The analysis of concurrence in the objects can be performed considering for instance the whole time frame and all objects. Therefore, a measure of concurrence is used. This measure indicates numerically how much the motion parameter values of one object correspond to the motion parameter values of all the other objects. This measure is computed for every object every time step and accumulated over the objects. Plotted in a *cross-classified table of the accumulated concurrence* (cumulative concurrence of object A to all others, B to all others, etc.) measures may give a condensed insight in the dynamics in the motion of the group.

Even though the Porcupine Caribou data set (Fig. 3) has almost the optimal data structure required for the REMO analysis concept, it has some structural shortcomings and thus illustrates some methodological challenges for the REMO analysis concept. First of all, the individuals are not localized absolutely synchronously. Even though the observation week numbers correspond, the exact day of observation may vary up to several days. Furthermore, the tracking interval varies irregularly over the duration of observation project from twice a week over weekly to every second week. As a straightforward approach to preprocessing, the REMO analysis concept works with a linear interpolation of the observation fixes in weekly intervals along the time axis. Last but not least some caribou died during the observation period (Gus-gus and Homer) and were replaced by other collared animals (Lynetta).

In the seriation procedure [9] the knowledge discovery in the analysis process happens by human cognition after reordering rows and columns in a matrix. In contrast, the REMO concept is designed to find patterns automatically and numerically without reordering the objects in the matrix. The STAM/STAC data mining approach [10] is made for localizing clusters in spatio-temporal point data, in basically immobile observation points of disease cases. In contrast, the REMO concept is designed for tracking data describing moving points. Whereas the time plot family [7] is designed to analyze the motion of one or at most two individuals, the REMO concept is designed for an unlimited number of objects.

5 Conclusions and Outlook

This paper presents an analysis concept for spatio-temporal observation data, called the REMO analysis concept. The REMO concept allows analyzing the relative motion of many moving point objects. The concept bases on the combination of two key ideas. First, several parameters describing the individuals motion are arranged in an analysis matrix. The rows of the matrix represent the individuals, the columns consecutive time steps. Second, spatio–temporal "behavior" and interrelations within groups of moving point objects are manifested as patterns in the REMO matrix. Basic patterns of relative motion such as *constance*, *concurrence* or *group turn* can in fact be identified and localized in space–time. The testing of the REMO analysis concept both on typical GPS–tracking data and abstract socio–political data, showed its universally applicable layout. The REMO analysis concept helps to discover interrelations in any kind of observation data of moving points objects, no matter whether the data describe moving animals, moving people or any kind of moving entity in artificial spaces.

The next steps of this ongoing research project include the advancement of a prototype environment to test and enhance the REMO analysis concept. The goal of this prototype is the implementation of measures and algorithms to describe and identify the patterns in a quantitative way. Therefore, we will define a pattern definition language. In the testing phase we will test the REMO analysis concept with constructed artificial and additional real observation data sets. Furthermore we will investigate the influence of different motion parameter classifications on the pattern detection results. Last but not least we will extend the system to extract arbitrary pattern inherent to the data.

Acknowledgments

We would like to thank the principal sponsors of the Porcupine Caribou Project for their support: ADFG, Yukon Renewable Resources, US Fish and Wildlife Service, Parks Canada, Government of the Northwest Territories, Canadian Wildlife Service and three anonymous referees for their valuable comments on earlier version of the manuscript.

References

1. Bertin, J. La Graphique et le Traitement Graphique de l'Information. Flammarion, Paris (1977)
2. Frank, A. Socio-Economic Units: Their Life and Motion. In: Frank, A., Raper, J., and Cheylan, J.-P. (eds.) *Life and Motion of Socio-Economic Units*. Taylor & Francis, London (2001) 21-34
3. Haggett, P. *Geography: A Modern Synthesis*. Harper and Row, New York (1983)
4. Hermann, M. and Leuthold, H. Weltanschauung und ihre soziale Basis im Spiegel der Volksabstimmungen. *Swiss Political Science Review* **7**:4 (2001) 39-63
5. Hornsby, K. and Egenhofer, M. Identity-Based Change: A Foundation for Spatio-Temporal Knowledge Representation. *International Journal of Geographical Information Science* **14**:3 (2000) 207-224
6. Hulbert, I. GPS and Its Use in Animal Telemetry: The Next Five Years. In: Sibbald, A. and Gordon, I. (eds.) *Proceedings of the Conference Tracking Animals with GPS*, Aberdeen, (2001) 51-60
7. Imfeld, S. *Time, Points and Space—Towards a Better Analysis of Wildlife Data in GIS*. Ph.D. Thesis, Department. of Geography, University of Zürich, Switzerland (2000)
8. Langran, G. *Time in Geographic Information Systems*. Taylor & Francis, London (1992)
9. Muller, J.-C. A Geoprocessing Tool for the Management of Spatial Data: Automated Seriation. In: Marble, D., Brassel, K., Peuquet, D., and Kishimoto, H, (eds.) *Proceedings of the International Symposium on Spatial Data Handling*, 1984, Zurich (1984) 127-139
10. Openshaw, S., Turton, I., and Macgill J. Using the Geographical Analysis Machine to Analyze Limiting Long-term Illness Census Data. *Geographical & Environmental Modelling*, **3**:1 (1999) 83-99
11. Peuquet, D. Making Space for Time: Issues in Space–Time Data Representation. *GeoInformatica*, **5**:1 (2001) 11-32
12. Wolfson, O, Xu, B., Chamberlain, S., and Jiang, L. Moving Object Databases: Issues and Solutions. In: *Proceedings of the 10th Int. Conference on Scientific and Statistical Database Management (SSDBM98)*, Capri, Italy (1998) 111-122

International Journal of Geographical Information Science
Vol. 19, No. 6, July 2005, 639–668

Taylor & Francis
Taylor & Francis Group

Research Article

Discovering relative motion patterns in groups of moving point objects

PATRICK LAUBE*, STEPHAN IMFELD and ROBERT WEIBEL
Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

(*Received 3 May 2004; accepted 7 January 2005*)

Technological advances in position-aware devices are leading to a wealth of data documenting motion. The integration of spatio-temporal data-mining techniques in GIScience is an important research field to overcome the limitations of static Geographic Information Systems with respect to the emerging volumes of data describing dynamics. This paper presents a generic geographic knowledge discovery approach for exploring the motion of moving point objects, the prime modelling construct to represent GPS tracked animals, people, or vehicles. The approach is based on the concept of geospatial lifelines and presents a formalism for describing different types of lifeline patterns that are generalizable for many application domains. Such lifeline patterns allow the identification and quantification of remarkable individual motion behaviour, events of distinct group motion behaviour, so as to relate the motion of individuals to groups. An application prototype featuring novel data-mining algorithms has been implemented and tested with two case studies: tracked soccer players and data points representing political entities moving in an abstract ideological space. In both case studies, a set of non-trivial and meaningful motion patterns could be identified, for instance highlighting the characteristic 'offside trap' behaviour in the first case and identifying trendsetting districts anticipating a political transformation in the latter case.

Keywords: Moving point objects; Geographic knowledge discovery; Data mining; Pattern matching; Temporal granularity

1. Introduction

Moving point objects (MPOs) are a frequent representation for a wide and diverse range of phenomena: for example, animals in habitat and migration studies (e.g. Ganskopp 2001, Sibbald *et al.* 2001), vehicles in fleet management (e.g. Miller and Wu 2000), agents simulating people for modelling crowd behaviour (e.g. Batty *et al.* 2003) and even tracked soccer players on a football pitch (e.g. Iwase and Saito 2002). All those MPOs share motions that can be represented as *geospatial lifelines*: a series of observations consisting of a triple of *id*, *location* and *time* (Mark 1998, Hornsby and Egenhofer 2002).

Gathering tracking data of individuals has become much easier nowadays due to substantial technological advances in position-aware devices such as GPS receivers, navigation systems and mobile phones. The increasing number of such devices is already leading to a wealth of data on space–time trajectories documenting the

*Corresponding author. Email: plaube@geo.unizh.ch

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810500105572

spatio-temporal behaviour of animals, vehicles and people for off-line, and potentially even on-line, analysis. These collections of geospatial lifelines represent a rich substrate to analyse individual and group behaviour. Understanding motion processes is an important precondition for the design of location based services (LBS) and traffic-control systems (Mountain and Raper 2001, Smyth 2001, Miller 2003). The database research and of course the GIScience communities are both challenged to provide methods to analyse and understand motion data. A short review of the database and GIScience body of literature highlights accomplished and open tasks exploring the motion of point objects.

1.1 *Database research*

In computer science, much work has been carried out studying the design, implementation, and application of spatio-temporal database management systems (STDBMS). These efforts contributed substantially to our knowledge of how to handle motion data in information systems. Joint research initiatives such as the CHOROCHRONOS project integrated knowledge about previously separated spatial and temporal databases to a spatio-temporal perspective: research has been carried out to build an ontology for STDBMS, trying to find the minimal commitment on how space and time is structured (e.g. Frank 2001, 2003). Data models and languages have been developed to express spatio-temporal relations and to design and query spatio-temporal databases (e.g. Grumbach *et al.* 2001, Erwig and Schneider 2002, 2003, Grumbach *et al.* 2003, Gueting *et al.* 2003). With regard to exploring motion, we focus here on the query functionality provided by STDBMS. Such spatio-temporal queries may involve data about moving entities. One could, for instance, query an air-traffic control database: 'Were any two planes close to a collision?' (Erwig *et al.* 1999, p. 271).

Traditional DBMS assume that data are constant until they are explicitly changed. Such databases are not well suited to handle the continuously changing positions of MPOs in real time, required for instance for a real-time taxi-cab fleet management system. Moving Objects Databases (MOD) feature data models with dynamic attributes, i.e. attributes that change continuously as a function of time, without being explicitly updated. Such a MOD allows queries like 'retrieve all airplanes that will come within 30 miles of the airport in the next 10 minutes' using a temporal query language, for example Future Temporal Logic (FTL) (Sistla *et al.* 1997, Wolfson *et al.* 1998, Trajcevski *et al.* 2004).

However sophisticated, the basic task of a query is to retrieve stored objects, collections of objects or their observations from a database. By contrast, analysis must go beyond querying and requires the production of new information and knowledge that is not directly observed in the stored data (Aronoff 1989, Golledge 2002). Whereas a lot of work has been accomplished in the OLTP (on-line transaction processing) area, only a few contributions can be found in the OLAP (on-line analytical processing) area, trying to investigate complex relationships and look for patterns, trends, and exceptions (see, for example, Rivest *et al.* (2001) and Bédard *et al.* (2003)). Hence, sketching the years ahead after the CHOROCHRONOS project, Koubarakis *et al.* (2003, p. 346) conclude: 'Of particular interest is here the mining of spatio-temporal patterns, since it can lead to important observations in many applications (e.g. environmental monitoring and fleet management). There is very little work in this area and much remains to be done.'

1.2 *GIScience*

Most GIS are based on a static place-based perspective and are thus still notoriously weak in providing tools for handling the temporal dimensions of geographic information (Mark 2003). Some GIScience approaches to analyse motion integrate time using the concepts of Hägerstrand's time geography with its space–time prisms (e.g. Miller 1991, Huisman and Forer 1998, Miller and Wu 2000, Hornsby and Egenhofer 2002, Kraak and Koussoulakou 2004). The space–time prism is a useful concept to analyse the motion constraints of a few individuals but is limited to very small numbers of individuals. Miller postulates expanding GIS from the place-based perspective to encompass a people-based perspective. Therefore, he identifies the development of a formal representational theory for dynamic spatial objects and of new spatio-temporal data mining and exploratory visualization techniques as key research issues for GIScience (Miller 2003).

Whereas the early days of Hägerstrand's time geography were limited to a data-poor and computation-poor environment, nowadays spatio-temporal analysis exists in data-rich and computation-rich environments. Knowledge discovery in databases (KDD) and data mining are reasonable responses to the huge data volumes in operational and scientific databases. 'KDD has evolved from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for experts systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in the context of large data sets' (Fayyad *et al.* 1996, p. 39). Data mining is just one central component of the overall KDD process. *Data mining* concerns the application of specific algorithms for extracting patterns from data. The central belief of KDD is that information is hidden in very large databases in the form of interesting *patterns* (Miller and Han 2001).

Whereas KDD and data mining have a long tradition in databases (Abraham and Roddick 1999, Roddick *et al.* 2001), these approaches are only starting to enter the field of GIScience. 'Traditional spatial analysis tools are far from adequate for handling the huge volumes of data and the growing complexity of analysis tasks. Geographic data mining and geographic knowledge discovery (GKD) represent an important direction in the development of a new generation of spatial analysis tools in data-rich environments' (Miller and Han 2001, p. 27). This statement is equally true for the spatio-temporal analysis of geospatial lifelines and is thus a key motivator for this research.

Our research seeks to contribute to the analytical power of GIScience to explore motion represented in geospatial lifelines. In this paper, work is presented which refines and extends a previously developed concept for GKD in motion data of MPOs (Laube and Imfeld 2002, Laube *et al.* 2004). The approach allows the formalization and identification of generic motion patterns in tracking data and extracting instances of these formalized patterns.

The remainder of this paper is structured as follows. Section 2 gives a short review of the RElative MOtion (REMO) analysis concept. Section 3 proposes an object-oriented approach to model and analyse the motion of many MPOs, and places special emphasis on dealing with varying temporal granularities in the analysis process. Section 4 introduces a formalism to describe and match motion patterns on geospatial lifelines. Section 5 illustrates an application prototype that implements the REMO analysis concept. Section 6 reports on experiments with two case studies. Section 7 discusses the concept, compares it with related approaches, and identifies strengths and open problems. Finally, conclusions are presented in section 8.

2. The REMO analysis concept

The key idea of the REMO analysis concept is to compare the motion attributes of point objects over space and time, and thus to *relate* one object's motion to the motion of all others (Laube and Imfeld 2002). Suitable geospatial lifeline data consist of a set of MPOs, each featuring a list of fixes. The REMO concept is based on two key features: First, a transformation of the lifeline data to a REMO matrix featuring motion attributes (i.e. speed, change of speed or motion azimuth). Second, formalized patterns are matched on this REMO matrix (figure 1).

Two simple examples shall illustrate the REMO analysis concept: Let the geospatial lifelines in figure 1(*a*) be the tracks of four GPS tracked deer. Deer O_1 is moving with a constant motion azimuth of 45° during an interval t_2 to t_5, i.e. four discrete time steps of length Δt (figure 1(*b*) and 1(*c*)). It is showing a *constancy* (figure 1(*d*)). By contrast, four deer performing a motion azimuth of 45° contemporaneously at t_4 show *concurrence*.

These two examples illustrate the two axes of the REMO matrix. Rows represent objects, columns represent time steps. A REMO pattern is defined as a set of motion parameter values with an extent in time (rows) and/or across the objects (columns). Patterns within one dimension of the matrix are the *simple* primitives, serving as a basis to compose *complex* patterns over time and across objects. The following is a subset of the simple REMO patterns introduced in Laube and Imfeld (2002, p. 136):

- *Constancy:* Sequence of equal motion attributes for r consecutive time steps (e.g. deer O_1 with motion azimuth 45° from t_2 to t_5)
- *Concurrence:* Incident of n MPOs showing the same motion attributes value at time t (e.g. deer O_1, O_2, O_3 and O_4 with motion azimuth 45° at t_4)
- *Change:* Change in an MPO's motion attributes of value v over r time steps (e.g. deer O_4 changes its motion azimuth from 90° to 0° during t_3 to t_5)

The consideration of interrelations not only in one, but in both dimensions of the matrix describes complex interactions between the motion of many objects. Complex patterns are in fact combinations of two or more simple patterns. The following example illustrates the composition of complex patterns. Combining

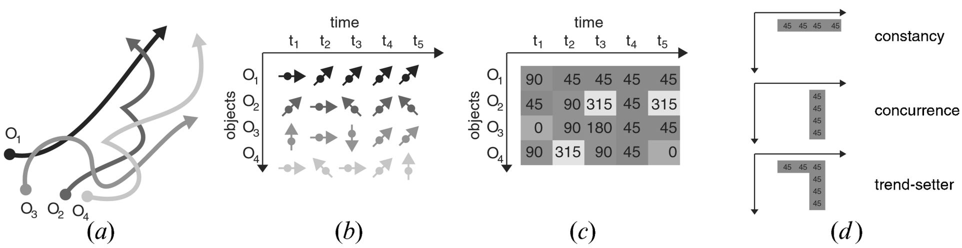


Figure 1. Geospatial lifelines of four MPOs (*a*) used to derive at regular intervals the motion azimuth (*b*). In the REMO analysis matrix consisting of classified motion attribute values (*c*) generic motion patterns such as *constancy, concurrence* or *trend-setter* are matched (*d*).

simple constancy with simple concurrence results in the complex pattern *trend-setter*:

- *Trend-setter:* One trend-setting MPO anticipates the motion of n others. Thus, a trend-setter pattern consists of a *constancy* linked to a *concurrence* (e.g. deer O_1 anticipates at t_2 the motion azimuth 45° that is reproduced by all other MPOs at time t_4)

For simplicity, we focus in the remainder of this paper on the motion attribute *azimuth*, even though most (but not all) facets of the REMO concept are equally valid for speed or change of speed as well as other motion attributes that might be determined.

3. Modelling motion attributes of MPOs

The REMO analysis concept deals with MPOs and events in their geospatial lifelines. The object-based nature of MPOs, events and instances of patterns is evident. Thus, the REMO world is an object-oriented world, and so is its class design.

3.1 *Modelling continuous and discrete motion*

The track of an MPO in the real world is always a continuous line. However, in order to record and analyse the pattern of its movement, the essential characteristics of an observed track have to be represented in a discrete form suitable for storage and analysis in an information system. Behavioural biologists, for instance, have spent decades of field work capturing the motion paths of diverse species such as butterflies, ants, birds, or red deer. They found that tracks can have different shapes depending on species-specific patterns of motion. The motion of some organisms is punctuated with periodic halts, e.g. flying butterflies periodically stop on vegetation. Although the actual track between two flowers may be circuitous, straight segments representing *discrete* moves appropriately summarize the displacements (Root and Kareiva 1984, Turchin 1991). Other organisms like ants or deer appear to move in a *continuous* way, not offering the analyst convenient stopping points. However, be it the sampling procedure of the tracking device or the need to translate the motion in a form suitable for storage and analysis, even continuous paths are usually discretized and approximated by a series of straight-line segments (Turchin 1998).

'Although most physical theories are spatially and temporally continuous the evaluation of these theories almost always requires the existence of entities which are spatially and temporally discrete' (Raper and Livingstone 1995, p. 362). Raper and Livingstone call this discretization an 'entification at a particular granularity'. Whereas space is usually discretized using the vector or raster models, time is normally discretized as a series of events or recordings. Discretizing continuous tracks, we can again benefit from the experience of the biologists in order to avoid undersampling and oversampling. If the track is sampled at a sampling granularity that is too coarse (undersampling), important information is lost. If, in contrast, we introduce excessive fixes (oversampling), the track's signal tends to be drowned out by noise, and feigned autocorrelation between successive moves is introduced. Hence, avoiding undersampling collecting data at the highest granularity possible is a simple yet effective strategy. The problem of oversampling is more subtle. One strategy is to resample the tracks at an increasingly coarse granularity, until autocorrelation between the moves disappears (see Turchin (1998, p. 130) for details on how to avoid oversampling when discretizing tracks).

The question of how to model motion also arises when it comes to designing and querying databases suitable for storing motion data (Gueting *et al.* 2003). There are two levels of abstraction in the design of database schemata for MPOs: modelling an MPO as acontinuous line would be a clean and simple *abstract model*. When it comes to implementation, the abstract model has to be translated into a *discrete model* that makes particular choices and thereby restricts the range of values of the abstract model that can be represented. The discrete model of an MPO would be a discrete set of segments. Or, to quote Erwig *et al.* (1999, p. 282): 'Abstract models are simple, but only discrete models can be implemented.'

3.2 *Imperfect data and temporal granularity*

The representation of the continuously changing position of an MPO is inherently associated with uncertainty (Pfoser and Jensen 1999, Cheng *et al.* 2004). For instance, how can the gaps between known positions be filled (Wentz *et al.* 2003)? The problem increases if the fixes are not sampled regularly but irregularly with changing time intervals and perhaps even missing values—a typical property of data originating from position aware devices. Thus, the MPO class design should offer ways of handling gaps to produce regularly sampled analytical derivatives such as a REMO matrix.

Whereas the term resolution refers to the least detectable difference in a measurement, analysis granularity stands for the level of detail, selectable by the user. Thus, the finest granularity need not be the most adequate for the analysis of a phenomenon. Consequently, the ability to change the temporal granularity, referred to as temporal zooming, is an important requirement for many scientific questions (Hornsby 2001). For spatio-temporal analysis and discovery, tools capable of changing between more and less detailed views are essential to solve a problem or to uncover information (e.g. Bettini *et al.* 2000). Support of temporal zooming is an important requirement on a class design for MPOs. Section 3.3 proposes a generic approach to derive motion attributes from irregularly sampled lifelines at variable granularities in order to allow temporal zooming in the REMO analysis concept.

3.3 *MPO class design*

In order to represent the motion attributes and behaviours (i.e. computational methods) of MPOs, an object-oriented design was chosen. The most prominent characteristic of the developed REMO class design is the strict separation of the MPO MODELLING DOMAIN and the ANALYSIS DOMAIN (figure 2). This separation allows the original tracking data to be managed separately from the analysis task. Thus, MPOs always keep their exact lifeline data and compute their motion attributes only on request for the analysis phase.

The class `MPO` maintains a list of `Fixes`, each holding a location `Point`. Based on these data, the MPO answers any spatio-temporal request on its (motion) attributes. The `REMOMatrix` class requests the motion attributes of the MPOs in accordance with its parameters, i.e. granularity and temporal extent of the matrix. `REMOPatterns` search their realizations (matches) on the `REMOMatrix`.

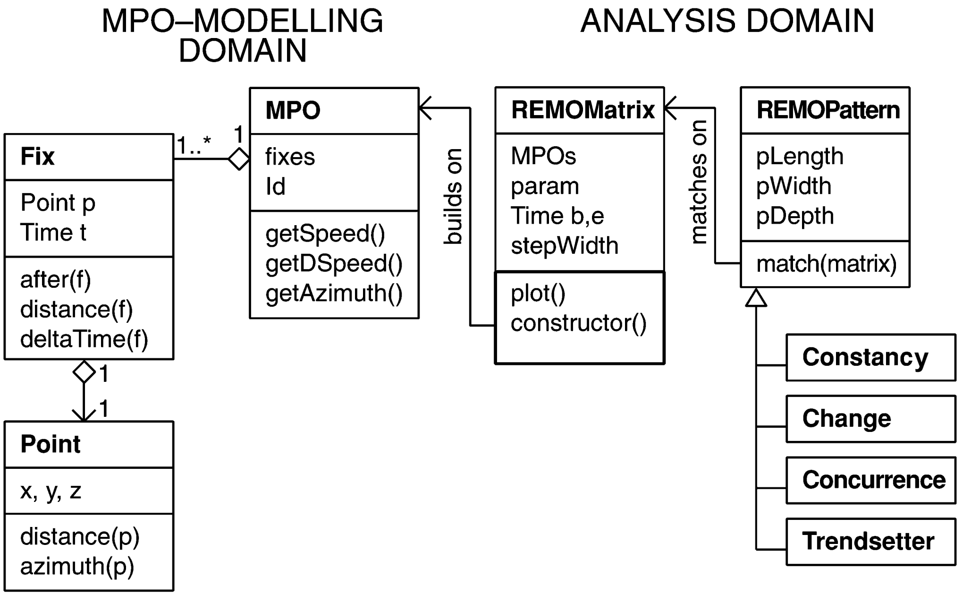


Figure 2. UML class diagram of the REMO class design showing the strict separation of the MPO MODELLING DOMAIN and the ANALYSIS DOMAIN.

Figure 3 shows the cooperation of the MPO MODELLING DOMAIN with the ANALYSIS DOMAIN. The `REMOMatrix` requests the motion attributes at its specified temporal granularity. For a finer granularity, the `REMOMatrix` requests more motion attribute values; hence, the involved `MPOs` may have to interpolate between fixes. For a coarser granularity, the `REMOMatrix` requests only few motion attribute values; hence, `MPOs` may have to aggregate motion attributes. The imperative of handling irregularly sampled fixes and variable temporal granularities therefore implies two problems: (1) interpolation of motion attributes in the case of irregular fixes or gaps of missing data; (2) aggregation of motion attributes in the case of a coarse analysis granularity over a much finer fixing rate.

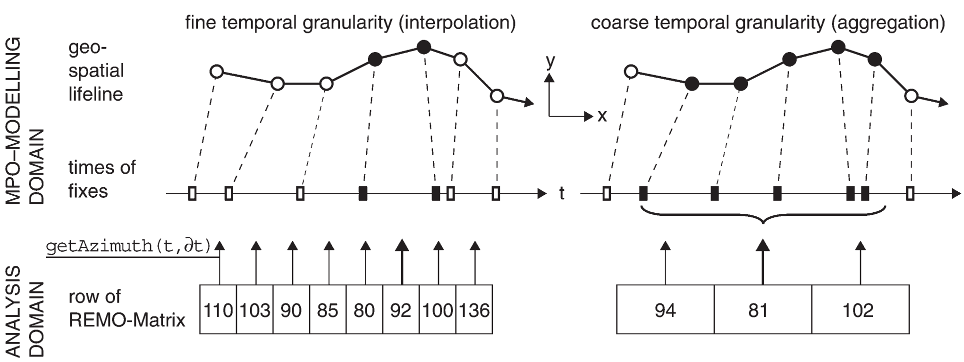


Figure 3. MPOs featuring detached attribute functions to derive motion attributes from the known lifelines at arbitrary query times (e.g. `getAzimuth(t,δt)`). Fine temporal granularities require interpolation between known fixes (e.g. the two fixes in black right before and after the query time), coarse temporal granularities require aggregation using a moving temporal window (e.g. the set of fixes in black within the given time interval).

The REMO class design resolves these issues with *detached attribute functions*. `MPOs` offer *attribute functions* to describe their motion based on their lifeline data (e.g. `getAzimuth`()). This can happen at any desired query time and with any desired granularity; the function is *detached* from the actual fixes. Thus, the functionality for temporal and/or spatial interpolation or aggregation of motion attributes and associated questions of fix uncertainty is encapsulated in the MPOs. This class design allows the integration of various detached attribute functions, using nearest-neighbour functions, focal functions or curve-fitting approaches. Nearest-neighbour functions compute the motion attributes considering simply the two nearest-neighbouring fixes of query time t. The focal functions use a moving window of length Δt around query time t to smooth inaccurate or incomplete data. Yet another approach is to first fit a smooth curve to the possibly scattered fixes and derive motion attributes from this smoothed lifeline. In short, the MPO MODELLING DOMAIN defines *how* to interpolate or aggregate motion attributes, and the ANALYSIS DOMAIN defines *when* this happens (figure 3).

4. Pattern description formalism

The REMO analysis concept follows the syntactic pattern detection approach in order to describe and detect predefined motion patterns in the geospatial lifelines of MPOs. After Jain *et al.* (2000), syntactic pattern recognition adopts a hierarchical perspective where a pattern is viewed as being composed of simple subpatterns, the primitives. Complex patterns are represented in terms of interrelationships between primitives. A formal analogy can be drawn between the structure of patterns and the syntax of a language. The patterns are viewed as sentences belonging to a language, primitives are viewed as the alphabet of the language, and the sentences are generated according to a grammar. Thus, in principle, any arbitrarily complex pattern can be described by a set of primitives and grammatical rules (Jain *et al.* 2000).

4.1 *Scope of the formalism*

The REMO analysis concept is designed to be a flexible and intuitive tool for researchers. Thus, it is an important precondition that users can compose patterns in a simple and flexible way. To allow flexible knowledge discovery, the patterns need parameters and descriptors to adjust their size and shape. Not only the basic motion patterns described in section 2 but also user-defined arbitrary patterns shall be formalised. Therefore, the REMO analysis concept comes along with a pattern description formalism to describe the introduced motion patterns. This formalism allows the potential users to compose REMO motion patterns in a simple, yet precise, compact and stringent way. As will be shown in the implementation section, combining the REMO formalism with a graphical user interface to compose motion patterns results in a effective framework for GKD. The proposed formalism is related to the commonly used regular expression formalism and to mathematical logic.

4.2 *Using concepts of regular expressions*

The commonly used regular expressions (regex) are a way of describing a set of strings without having to list all the strings in the set (Wall *et al.* 1996). Regex are basically used to determine whether a string matches a particular pattern.

Commonly, regex are used to search, edit and manipulate string data (e.g. Wall *et al.* 1996, Friedl 2002). Regex are widespread in UNIX programs (grep), editors (emacs, vi) and programming languages (perl, java, Tcl, Python). As an example, the regular expression $a\{3,5\}$ matches the bold characters in the following string:

*bcbaabc***aaa***cbc***aaaaa***ccc*

There are three major differences between pattern matching in strings using regex and REMO pattern matching. First, and most obviously, the REMO analysis concept requires two-dimensional pattern matching on a two-dimensional matrix, whereas regex is normally used to match patterns on one-dimensional strings. Second, pattern matching on the REMO analysis matrix is confronted with two different types of dimensions. Whereas the temporal axis is an interval scale, and thus comparable to sequences of strings, the object axis implies no order among the objects. Thus, adjacency among objects is arbitrary (figure 1). Third, the elements constituting a pattern are numbers on a continuous scale, rather than characters and numerals as with regex. These three differences lead to the differences from basic regex in the formation of the REMO pattern-matching expressions.

Nevertheless, many features of regex are very similar to the requirements for a REMO pattern matching formalism. Whenever possible, the REMO formalism uses familiar regex structures to express REMO patterns. That applies, for instance, to the descriptors referring to the number of time steps and number of MPOs building a pattern (pattern length and width, respectively; see section 4.3.2). Slight changes had to be made to describe the range of attribute values building a pattern (pattern depth; see section 4.3.3). The most prominent changes emerged from the need to combine *simple patterns*, in either the temporal or the objects dimension, to build the *complex patterns* extending across dimensions, such as for a trend-setter.

4.3 *Structural elements of the REMO pattern formalism*

The most important principle of language design says that easy things should be easy, and hard things should be possible (Wall *et al.* 1996). This rule also applies for the development of the REMO formalism. The easy tasks are expressing the simple patterns akin to known regular expressions; the harder problems involve the formalization of the complex patterns over two matrix dimensions.

4.3.1 Simple patterns over time or across objects. First of all, the dimension of the simple patterns in the analysis matrix must be specified. A pattern P over time describes a sequence S of motion attribute observations A_{m} (expression (1)). A pattern P across objects describes an incident I of a set of concurrent motion attribute observations A_{m} (expression (2)).

$$P = S(A_{\mathrm{m}}) \tag{1}$$

$$P = I(A_{\mathrm{m}}) \tag{2}$$

The REMO formalism allows the formalization of identical patterns in different but synonymous forms. Quantifiers allow A_{m} to be specified. Attribute values v are given in brackets. REMO patterns span time and across objects, and their extent is expressed with quantifiers in braces. Patterns over time have a length l, and patterns across objects have a width n (expressions (3) and (4)).

$$P = S(v_1, v_2, v_3, \ldots, v_l) = S([v]\{l\}) \tag{3}$$

$$P = I(v_1, v_2, v_3, \ldots, v_n) = I([v]\{n\}) \tag{4}$$

A change C is a special form of a pattern over time. It is formalized using an indicator for a starting attribute value v in brackets, for the value of change δv and a length l in braces (expression (5)). The starting attribute value indicator $[v]$ is optional, since a starting attribute value is not always desired.

$$P = C([v]\{\delta v\}\{l\}) \tag{5}$$

Change may be directional, and some attribute sets are cyclically closed (e.g. the compass rose for the motion azimuth). Increasing change is indicated using $+\delta v$, decreasing change using $-\delta v$ (expressions (6) and (7)).

$$P = C([v]\{+\delta v\}\{l\}) \tag{6}$$

$$P = C([v]\{-\delta v\}\{l\}) \tag{7}$$

Examples: A single deer heading north-east for four consecutive time steps is formalized as $P=S(45, 45, 45, 45)$ or $P=S([45]\{4\})$ (case a in figure 4). By contrast, four deer all heading north-east at the same time are formalized as $P=I(45, 45, 45,$

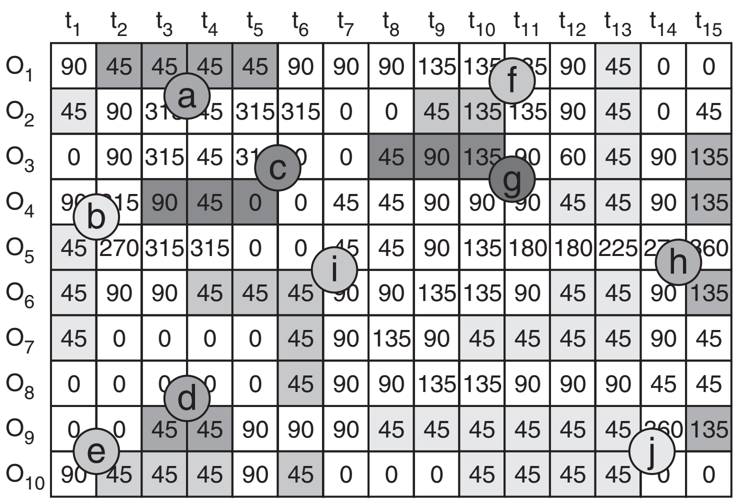


Figure 4. Hypothetical REMO matrix for 10 deer over 15 time steps. REMO formalism must be able to describe the basic and arbitrary REMO patterns. A set of example REMO patterns is highlighted: constancy (*a*, *d*, *e*), concurrence (*b*, *h*), trend-setter (*i*), change (*c*, *f*, *g*), and infection (*j*).

Table 1. Quantifiers.

	Description
Length quantifiers	
$\{l\}$	Exactly l time steps (e.g. 4 in a row)
$\{l,\}$	At least l time steps (e.g. at least 4 in a row)
$\{l, k\}$	Between l and k time steps (e.g. between 4 and 6 in a row)
Width quantifiers	
$\{n\}$	Exactly n objects at same time (e.g. 4 objects)
$\{n,\}$	At least n objects at same time (e.g. at least 4 objects)
$\{n, m\}$	Between n and m objects (e.g. between 4 and 6 objects)
Element quantifiers	
$v?$	Motion attribute value v is optional
v^*	0 or more times motion attribute value v
$v+$	1 or more times motion attribute value v

45) or $P=I([45]\{4\})$ (case b in figure 4). A motion azimuth change of one deer switching from east (90°) to north (0°) within three time steps would be formalized as $P=C([90]\{-90\}\{3\})$ (case c in figure 4).

4.3.2 Quantifiers. Much like the single characters in strings do in regex, REMO pattern elements can have quantifiers in the REMO formalism (table 1). In our case, the quantifiers are used to describe the pattern extent in the temporal and the objects axis. The pattern *length* (time axis) can be stated as *fixed* $\{l\}$ (exactly l time steps), *open* $\{l,\}$ (at least l time steps) or as a *range* $\{l, k\}$ (between l and k time steps). The pattern *width* (object axis) is formalized in an analogous way: *fixed* $\{n\}$ (exactly n individuals), *open* $\{n,\}$ (at least n individuals) or as a *range* $\{n, m\}$ (between n and m individuals). Thus, a constancy pattern can either have a fixed, open or range length.

A question mark (?) makes the preceding pattern element optional. An asterisk (*) refers to a pattern element that may appear never or many times. A plus sign (+) describes pattern elements that appear once or many times.

Examples: Analysing lifelines of deer, one might look for individuals moving constantly towards north-east for a certain time period. Depending on the exact scientific question, this could be stated with a fixed pattern $P=S([45]\{3\})$ (case e in figure 4), with an open pattern $P=S([45]\{2,\})$ (cases a, d, and e), or with an range pattern $P=S([45]\{2, 3\})$ (cases d and e only).

As a further example, one might want to identify individuals heading first north-east (45°) and subsequently turning through east (90°) to south-east (135°). Quantifiers allow every single step of this motion to be stated more precisely. The expression $P=S(45, 90?, 135)$ makes the eastward motion step optional and matches cases d and e in figure 4. With $P=S(45, 90^*, 135)$, an arbitrary series (including 0 times) of eastward motion steps can be interposed, and again (d) and (e) match. $P=S(45, 90+, 135)$ requires at least one eastward motion step, which excludes (d) from the matches.

4.3.3 Pattern depth descriptors. Since the pattern elements are numbers and not characters or numerals, the REMO formalism has its own descriptors making it possible to express the pattern element's value ranges referred to as pattern *depth* descriptors (see table 2). Exact motion attribute values are expressed in brackets $[v]$. The pattern depth descriptor can take the common relational operators less than

Table 2. Pattern depth descriptors.

Depth quantifiers	Description
$[v]$	Exactly motion attribute value v (e.g. azimuth 45°)
$[\leqslant v]$	At least motion attribute value v (e.g. azimuth 45°or more)
$[\geqslant v]$	Not more than motion attribute value v (e.g. azimuth 45° or less)
$[>v]$	More than motion attribute value v (e.g. more than azimuth 45°)
$[<v]$	Less than motion attribute value v (e.g. less than azimuth 45°)
$[v-w]$	Motion attribute value range between v and w (e.g. azimuth between 45° and 135°)
$[\wedge v]$	All possible motion attribute values except v
$[u\|v\|w]$	Motion attribute values u or v or w (e.g. 45° or 90° or 135°)

($<$), greater than ($>$), less than or equal to ($\leqslant$), and greater than or equal to ($\geqslant$). Ranges are expressed with a hyphen ($-$), exclusion with a caret (^). The OR operator allows a choice of pattern elements to be specified, using a vertical bar (|). Since a pattern element is unique, the AND operator (&) is omitted.

Examples: One possible use of pattern depth descriptors is to find an individual deer that first moves exactly north-east and then moves on in a direction between north-east and south-east. This pattern is formalized as $P=S(45, [45–135])$. One match amongst many others in figure 4 is instance (f).

4.3.4 Unbound patterns. So far, all the patterns consisted of clearly specified motion attribute observations A_{m}. However, some users might not know in advance which exact values the REMO patterns in their data will have. There might, for instance, lurk constancy patterns of u as well as v or w in a data set. In this case, the constancy pattern has a shape but no defined content, and the pattern is *unbound*. A *bound* pattern in contrast has defined motion attribute observations. The REMO formalism features the wildcard [#] to express unbound patterns. Pattern P in expression (8) matches any sequence of l consecutive equal motion attribute observations. Pattern P in expression (9) matches any set of n concurrent motion attribute observations.

$$P=S([\#]\{l\}) \tag{8}$$

$$P=I([\#]\{n\}) \tag{9}$$

Examples: $P=I([\#]\{4\})$ matches any set of four deer concurrently moving in the same direction. Two possible instances of P are highlighted as (b) and (h) in figure 4. As will be seen in the implementation section, unbound pattern matching has been chosen to keep the prototype as generic as possible. Diverse filtering procedures turn unbound pattern matching back into bound pattern matching.

4.3.5 Complex patterns over time and across objects. The REMO formalism presented so far is closely related to basic regex. The major differences come with the need to express two-dimensional patterns on a two-dimensional analysis matrix. The solution chosen for the REMO formalism follows an intuitive and simple approach. Since complex patterns are defined as a composite of simple patterns, i.e. sequences and incidents, it is straightforward to reproduce this construction principle in the formalism. A complex pattern is formalized as a set of simple patterns that are temporally linked (expression (10)):

$$complexPattern_i = \begin{cases} simplePattern_1 & : & interval_1 \\ simplePattern_2 & : & interval_2 \\ simplePattern_3 & : & interval_3 \\ \cdots & : & \cdots \\ simplePattern_n & : & interval_n \end{cases} \tag{10}$$

The term $interval_i$ allows indication of the relative temporal order of the linked simple patterns. Although in many cases, a complex pattern will either have a common shared start time t_b or end time t_e, this is not required to build a complex pattern.

The following set shows three trend-setters with different lengths $\{l\}$, $\{l,\}$, and $\{l, k\}$, respectively, expressions (11), (12), and (13):

$$P = \begin{cases} S([v]\{l\}) & : & t_{e-l+1}, \ldots, t_e \\ I([v]\{n\}) & : & t_e \end{cases} \tag{11}$$

$$P = \begin{cases} S([v]\{l,\}) & : & t_{e-j}, \cdots, t_e | (l-1) \leqslant j \leqslant (e-1) \\ I([v]\{n\}) & : & t_e \end{cases} \tag{12}$$

$$P = \begin{cases} S([v]\{l, k\}) & : & t_{e-j}, \cdots, t_e | (l-1) \leqslant j \leqslant (k-1) \\ I([v]\{n\}) & : & t_e \end{cases} \tag{13}$$

Quantifiers can be used not only to describe sets of single REMO matrix cells but also to compose complex patterns. For example, the contemporary occurrence of n identical constancy patterns of length l could be viewed as a concurrence of n constancy patterns and thus be expressed as a nested term (14).

$$P = I(S([v]\{l\})\{n\}) \tag{14}$$

Examples: Investigating group dynamics in a herd of deer, one might search for an alpha individual initiating a travel in a north-east motion before all other members of the herd. Such a trend-setter pattern P is shown in case i figure 4. Deer O_6 anticipates at time t_4 three time steps in advance the motion of the deer O_7, O_8, and O_{10} (expression (15)):

$$P = \begin{cases} S([45]\{3\}) & : & t_{e-2}, \ldots, t_e \\ I([45]\{4\}) & : & t_e \end{cases} \tag{15}$$

A slight modification of the trend-setter pattern illustrates the potential of the REMO formalism. The pattern of interest shall be called *infection* (case j in figure 4): a deer starts to move along the motion azimuth 45° at t_b. After two time steps, two other deer join this emerging group and show the same motion. Another two deer join one time step later. A set of three deer are joining the group at t_{b+5} building an incident of width 8 (16):

$$P = \begin{cases} S([45]\{6\}) & : \quad t_b, \ \cdots, \ t_{b+5} \\ I(S([45]\{4\})\{2\}) & : \quad t_{b+2}, \ \cdots, \ t_{b+5} \\ I(S([45]\{2\})\{2\}) & : \quad t_{b+4}, \ \cdots, \ t_{b+5} \\ I([45]\{8\}) & : \quad t_{b+5} \end{cases} \tag{16}$$

The REMO formalism offers a simple and comprehensible way to describe arbitrary REMO patterns. The formalism is used to describe the pattern-matching process in the following sections.

5. Implementation

As a proof of concept, the REMO analysis approach was implemented as an application prototype in Java. Three reasons argued for a stand-alone solution instead of customizing an out-of-the-box GIS. First, there was the need for an object-oriented spatio-temporal class design for MPOs, not present in most of today's commercial GIS. Second, high demands on integrating the pattern-matching process with visualization required an open and flexible environment. Third, only a few traditional GIS functions were required.

To maximize the intuitiveness of use for potential users (e.g. wildlife biologists or social scientists), we refrained from developing an interpreter for the REMO formalism. Hence, we decided to develop an application prototype featuring an easy-to-use graphical user interface (GUI). Thus, the REMO prototype features modules to manage and preprocess the data, to animate the MPOs in a space–time viewer (figure 8), to control the pattern matching process (figure 5) and to visualize results.

5.1 *GKD process*

The example of matching the pattern $P = S([\#]\{4,\})$ illustrates the GKD process (figure 6). Upon user request, the `mainController` first calls `create` to build an `MPOGroup`. The `MPOGroup` calls subsequently `create` to construct a `matrix` based on the lifelines of the MPOs. The `matrix` is built according to the properties set up by the `mainController`: the dimensions, the temporal granularity

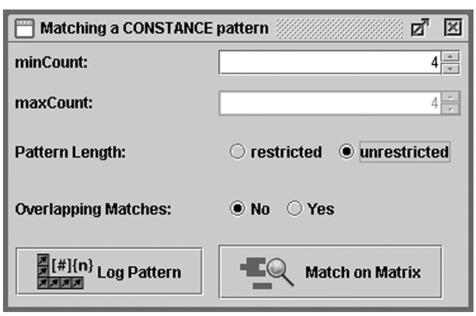


Figure 5. Pattern descriptors like pattern length, width and depth can be specified in different interfaces. The REMO pattern specified in this example is the constancy $P = S([\#]\{4,\})$.

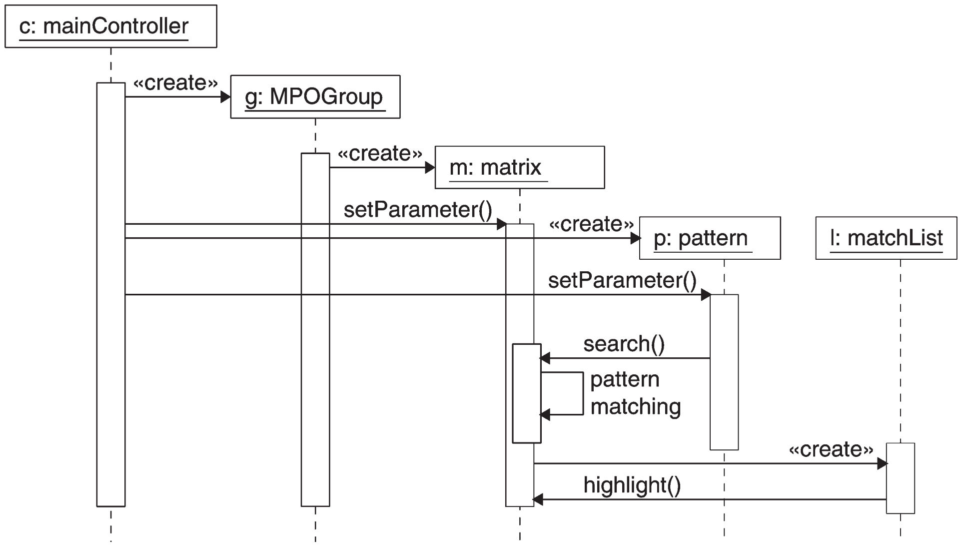


Figure 6. Simplified UML sequence diagram of the implemented REMO pattern-matching process.

and the motion attribute of interest and its reclassification (`setParameter()`). Under the control of the users, the `mainController` calls `create` to set up (`setParameter()`) a `pattern` to `search` on the `matrix`. The `pattern matching` process locates instances of the searched pattern on the `matrix`. A `matchList` gathers the resulting matches (`create`) for later visualization (`highlight()`) and further investigations. This process can be repeated with varied matrix and pattern properties.

Note that the pattern matching itself is two-tiered. First, unbound patterns like $P = S([\#]\{4,\})$ are constructed (figure 5) and matched on the matrix. Only in a downstream filtering process can users select patterns of specified values from the `matchList`. This step reduces in this example the matches of $P = S([\#]\{4,\})$ to the probably fewer matches of $P = S([45]\{4,\})$.

5.2 *Pattern-matching algorithms*

Matching REMO patterns on the REMO matrix has a certain similarity to the classical pattern matching problem on strings. Given a text string T of length n and a pattern string P of length m, one wants to know first whether P is a substring of T and second where on T this match is located. Decomposing a REMO matrix into its rows (motion attribute arrays) and columns (time-slices) allows the use of derivatives of classical string pattern-matching algorithms like the *Brute-Force Pattern Matching* (BFPM) or *Knuth–Morris–Pratt* (KMP) (Knuth *et al.* 1977). While pattern matching on strings compares characters, REMO pattern matching has to compare numbers which can be associated with relational operators. However, the basic principles remain the same. Complex patterns are matched by first matching constitutive simple patterns in either the temporal or object dimension and then testing whether the additional conditions over both dimensions are also met.

Brute-force pattern matching simply tests all the possible placements of P relative to T in the worst case in $O(nm)$ running time (Goodrich and Tamassia 1998). KMP uses a failure function f for the pattern string P which encodes repeated substrings inside the pattern itself to reuse previous comparisons and thus avoid unnecessary comparisons. It achieves in the worst case a running time of $O(n+m)$ (Goodrich and Tamassia 1998). The performance of the application prototype did not show worst-case behaviour and thus did not encounter significant performance problems with the above-mentioned algorithms and with the case-study data used so far. With larger data volumes, there may be a requirement to develop and use more sophisticated pattern-matching algorithms.

6. Case studies

The REMO analysis concept and its implementation prototype have been tested with various data ranging from tracked individuals to moving data points in abstract spaces. Two different case studies illustrate the concept's generic applicability: football (soccer) players tracked on the pitch and data points in an abstract ideological space.

6.1 *Football players*

From a non-scientific perspective, the motion of football players is probably the most intensively observed and most competently discussed motion of individuals ever created by human culture. However, from a GIScientist's perspective, a team of football players is a group of MPOs acting in a structured way on a well-defined space and over a well-defined time period. Thus, football players are an ideal case study to evaluate the REMO analysis concept. The motion of a team of football players is highly coordinated. Luring the other team into the offside trap, for example, requires coordinated motion of all four defenders in a row. Another example might be if a creative striker governs the motion of the row of defenders at the back.

The data used emerged from research pursued with the aim of tracking football players using multiple television cameras (Iwase and Saito 2002, 2003). The time frame covers about 33 of a football game in a Japanese University league, tracking 11 players with a sampling rate of 15 fixes per second (figure 7). Note the confusing tangle of lifelines produced by such a small group of 11 individuals and such a short time frame of half a minute.

For simplicity and illustration purposes, we focus on the motion azimuth in this example. The REMO matrix in figure 7 has a granularity of one fix per second to simplify the matrix and to eliminate short-term positional noise. Since the pitch's orientation in this example is left to right, the team is attacking to the east (azimuth 90°) and defending to the west (azimuth 270°).

First, we search for *constancy*. Constancy with regard to tracked football players describes players running in one direction for some period. This pattern can especially be expected for the left and right wingers as well as for the strikers. The former repeatedly sprint along the side-lines, the latter striking from the midfield in the direction of the opponent's goal. As a first simple example, we search for the longest straight tracks in this data sample. Figure 8 illustrates the matches for a constancy of at least length 10 ($P=S([270]\{10,\})$). The three players No. 5, 14,

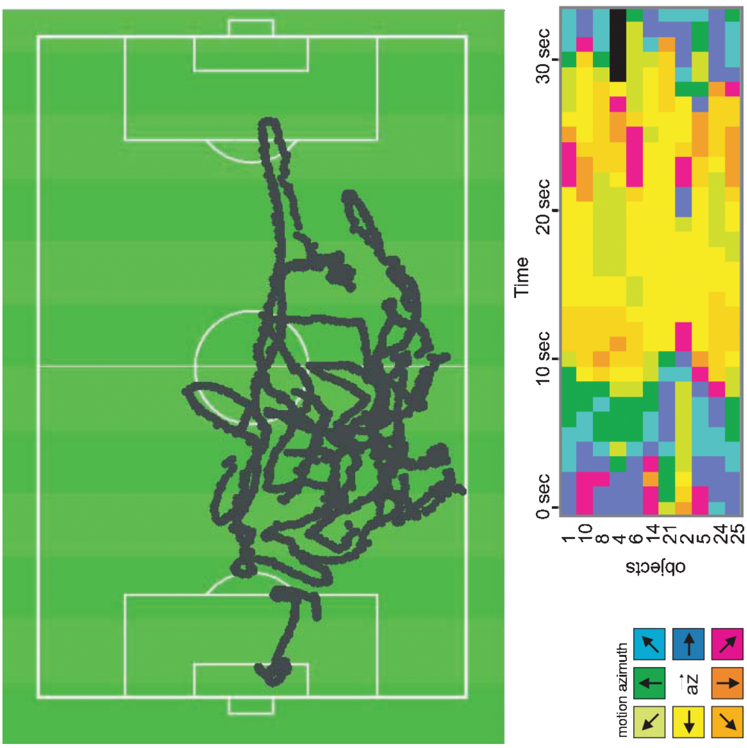


Figure 7. Football pitch showing the geospatial lifelines of 11 players covering a time frame of approximately 33 s. The REMO analysis matrix below illustrates the players' motion azimuth at a granularity of one fix per second. The most prominent group motion in this interval of the match is a backwards motion of the team establishing a proper defence formation around $t=15$ s.

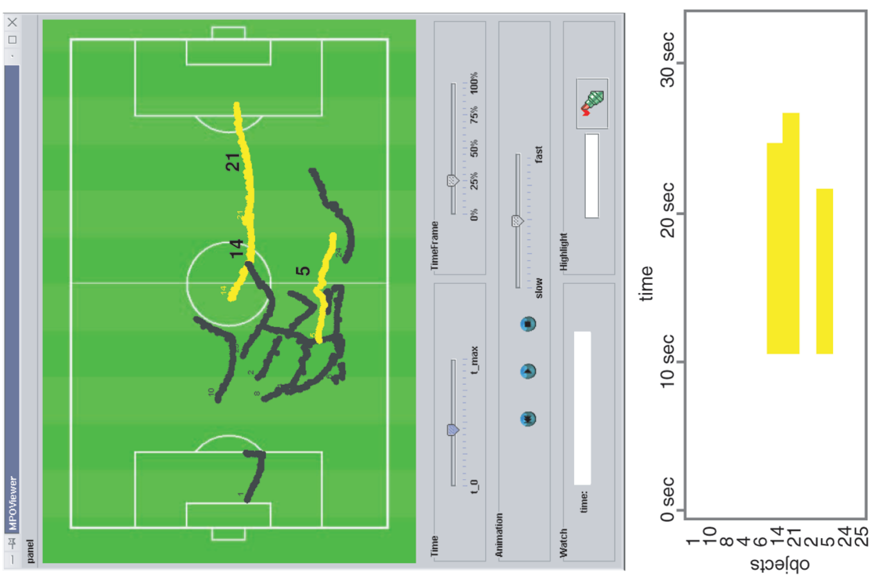


Figure 8. Three matches of a constant motion azimuth 270° over at least 10 s. Note that the motion azimuth is classified into eight discrete classes and computed over a temporal interval with a specific length. Hence, the tracks of the matched objects 5, 14, and 21 are considered to be straight to the east, even if the exact values may be scattering slightly around 270°.

and 21 showing this long straight sprint are indeed offensive players with the longest way back and thus forced to take the straight line.

As a second pattern, we try to identify a *concurrence*. Common sense and a first inspection of figure 7 suggest that concurrent motion is often seen in the lifelines of a team of football players. For example, the maintenance of an effective offside trap demands highly coordinated motion from the defenders. If we are interested in the degree of coordination in the players' motion, we can search for concurrence incidents with as many participants as possible. In the present example, a concurrence consisting of 10 out of 11 players can be found. The event shown in figure 9 illustrates the reaction of almost the whole team to an attack at time t_{15}; 10 MPOs move backwards, with an azimuth of 270°, respectively formalized as ($P=I([270]\{10,\})$).

From a football coach's perspective, *trend-setters* might be of special interest. A trend-setting football player might anticipate the important moves in the game. For this illustration of the concept trend-setter, we focus on the coordinated defending of the team around t_{14}. The general exploratory task is to find trend-setters anticipating this move. In the example in figure 10, we identify a trend-setter of at least length 4 and with at least eight team mates joining in the backwards move of the leading individual.

$$P=\begin{cases} S([270]\{4,\}) & : \quad t_{e-3},\ \ldots,\ t_e \\ I([270]\{8,\}) & : \quad t_e \end{cases} \tag{17}$$

Football scene analysis using multiple TV cameras has a huge potential for strategy understanding and making digest TV programs. Investigating the emerging vast amount of lifeline data from sports applications requires spatio-temporal data-mining methods. We identify this field as an opening opportunity to bring in the knowledge and tradition of GIScientists analysing spatio-temporal data.

6.2 *Abstract data points*

The frequently held popular referendums in Switzerland (approx. eight to 10 per year) allow researchers to make detailed inferences about value conflicts within the society. Hermann and Leuthold (2001) developed an inductive approach to discover the basic ideological conflicts in Switzerland. Performing a factor analysis on referendum data at the district level of all 158 federal referendums held between 1981 and 1999, they hypothesized a structure of mentality, which was interpreted as being composed of three dimensions: political left vs. political right, liberal vs. conservative and ecological vs. technocratic. The axes of this multidimensional ideological space, taken in pairs, provide a total of three two-dimensional maps of the political landscape of Switzerland (Hermann and Leuthold 2003). In these two-dimensional ideological spaces, the 185 districts can be localized in intervals of one year, from 1981 until 1999. Irrespective of their political and social meaning, the districts can be considered as moving points in a two-dimensional space (figure 11).

Figure 12(*a*) shows the REMO matrix for the motion azimuth of the districts. The districts of a Canton (member states of the Swiss Federation) are grouped together in the matrix (figure 12(*a*)). Thus, proximity in the matrix corresponds to a certain institutional and cultural similarity.

To show the generic potential of the REMO analysis concept, we investigate the old argument as to whether the German and the Latin part (French, Italian and

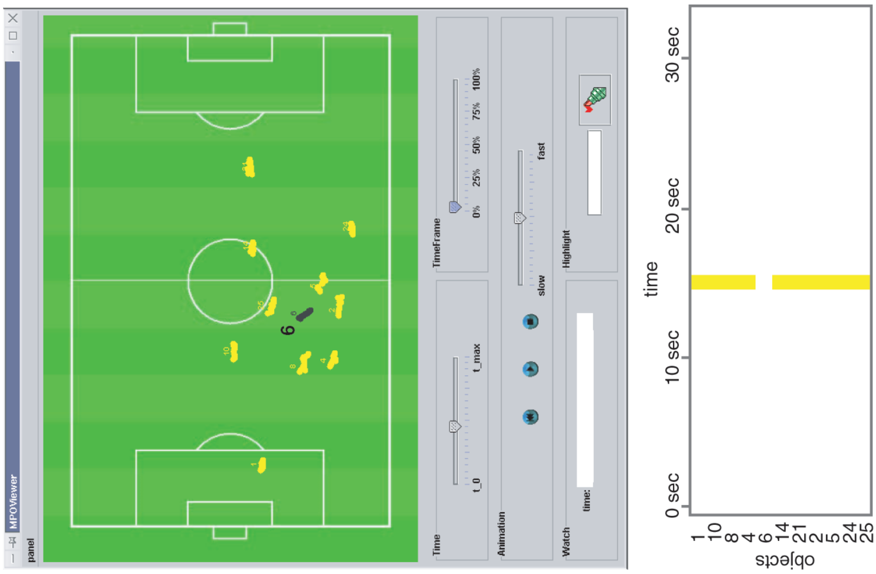


Figure 9. At t_{15}, almost the whole team synchronously moves back to the own goal to establish a defence formation. Only player 6 is classified to 315° instead of 270°.

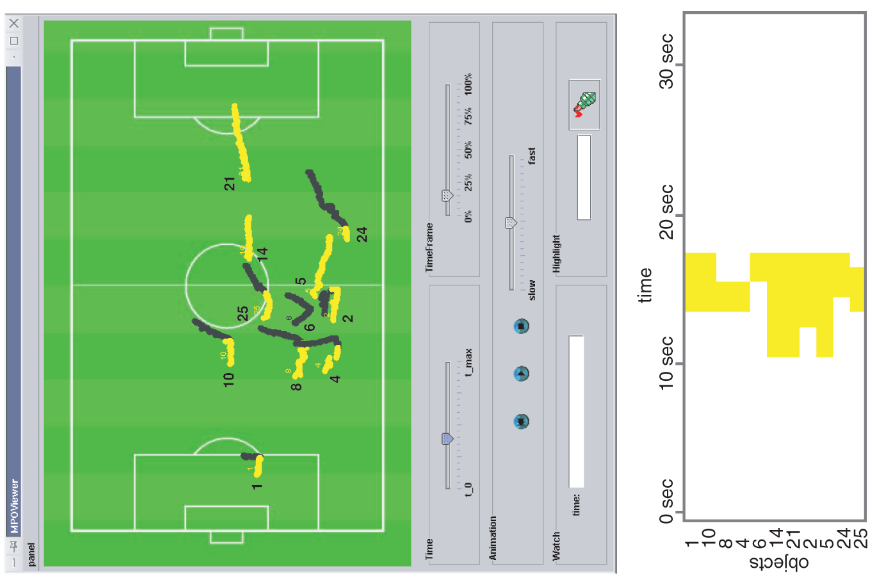


Figure 10. Lifelines in the football pitch illustrating the early backwards motion of players No. 5, 14, and 21. Only at t_{14} do the followers No. 1, 2, 4, 8, 10, 24, and 25 join in to complete the trendsetting pattern described in formula (17). Note that the REMO matrix plot contains 16 overlapping instances of the same trend-setter pattern.

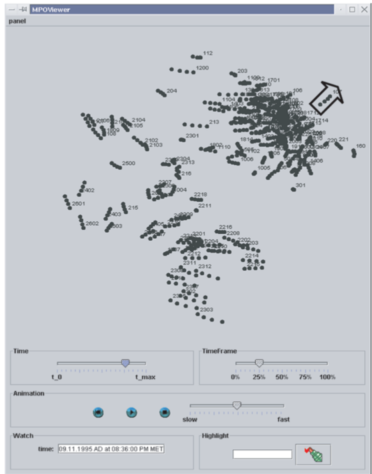


Figure 11. Motion of the 185 districts in the ideological space developed by Hermann and Leuthold (2003) using the MPO Viewer of the REMO prototype application. The uppermost frame contains the abstract ideological space spanning between the dimensions political left (left) vs. political right (right) and ecological (top) vs. technocratic (bottom). Every point chain holds the consecutive fixes of one district. For this screen shot, the time-frame controlling sliders at the bottom were set to the time *1995* and a time-frame length of *4* years. Thus, each point chain represents all fixes of its district in the ideological space between 1993 and 1997. The arrow in the top-right corner highlights the motion of the district Meilen to the 'ecological-right' corner of the ideological space.

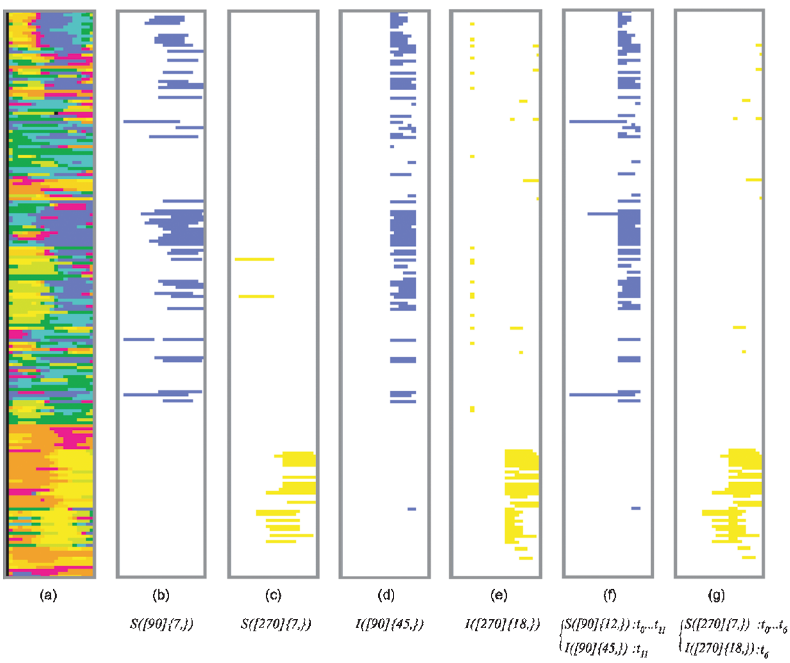


Figure 12. Political divergence of the German-speaking part from the francophone part of Switzerland in the 1990s. (*a*) Overview of the motion azimuths of the 185 districts from 1981 until 1999. (*b*, *c*) Constancy patterns towards the right respectively left political pole. (*d*, *e*) Concurrence patterns of the trend. (*f*, *g*) Trend-setters of the trend.

Rhaeto-romanic) of Switzerland are politically diverging, opening the so-called 'Röschtigraben' ('Röschti' being a characteristic Swiss-German food, and 'Graben' being a ridge; literally 'röschti-ridge', the concept used to describe the cultural divide between the two parts of country). Since we are not political scientists, it is not our intention to derive any statistically significant political conclusions. In fact, we merely want to demonstrate the exploratory potential of our approach, identifying segregation tendencies to motivate further political research on the topic. Such tendencies can be extracted with the REMO concept, detecting for constancy patterns to show that there really is a notable sum of districts moving in a specified direction. For this illustration, we searched for constancy patterns with a length of 7 years, a third of the entire period of 20 years.

Figure 12(*b*) reveals in the upper two thirds 45 districts showing $P=S([90]\{7,\})$ standing for a rightwards motion of the German speaking districts in the 1990s (36% of all German speaking districts). The lower third of the plot in figure 12(*c*) shows 18 constancies of $P=S([270]\{7,\})$ illustrating the opposite motion to the left by the francophone districts (39% of the francophone districts). As an interesting feature, two German-speaking districts could be identified that show a fairly isolated constant leftward drift in the 1980s (in the centre of figure 12(*c*), Sargans SG and Stein SH).

Having found so many constancies of length 7 in only one decade points to the assumption that there must be concurrences, representing the divergence of the German- and French-speaking districts. Those concurrences are expected to spread over a width corresponding approximately to the sum of the constancies found above. Indeed, figure 12(*d*) shows six subsequent concurrences $P=I([90]\{45,\})$ mainly in the German-speaking districts and figure 12(*e*) a striking over-representation of French speaking districts involved in one of the nine occurrences of $P=I([270]\{18,\})$.

Finally, the question arises as to whether the REMO analysis concept could identify the districts anticipating the left–right divergence. Figure 12(*f*) and (*g*) give the first hints on this issue, showing all matches to the following trend-setter pattern, again mapping approximately the dimensions of the constancies and concurrences found:

$$P=\begin{cases} S([90]\{7,\}) & : \quad t_0,\ \ldots,\ t_6 \\ I([90]\{45,\}) & : \quad t_6 \end{cases} \tag{18}$$

This trend-setter pattern can be matched a total of 72 times. The real trend-setting districts are those who anticipate the rightward drift the earliest. To identify them, the trend-setter pattern is expanded to:

$$P=\begin{cases} S([90]\{12,\}) & : \quad t_0,\ \ldots,\ t_{11} \\ I([90]\{45,\}) & : \quad t_{11} \end{cases} \tag{19}$$

Figure 12(*f*) shows the result. Three districts can very clearly be identified as the trend-setters (Trachselwald BE, Zofingen AG and Gösgen SO). The first two districts anticipate more than 10 years in advance a motion attribute that at least 45 followers later adopt. To identify trend-setters in the leftward drift of the French-speaking districts, the following pattern has been matched:

$$P = \begin{cases} S([270]\{7,\}) & : \quad t_0, \ldots, t_6 \\ I([270]\{18,\}) & : \quad t_6 \end{cases} \tag{20}$$

Figure 12(*g*) helps to identify the first two districts heading left (Conthey VS and Entremont VS).

The spatialization of statistical data, for example by plotting the annual shift of political entities in a ideological space, provides an opportunity for motion analysis. GIScience may contribute with its analysis and visualization potential. Both case studies underscore the need for designing generic methods for spatio-temporal knowledge discovery, as proposed with the REMO analysis concept.

7. Discussion

This discussion first evaluates our approach with respect to other approaches providing analysis tools for spatio-temporal data and their potential to analyse motion data. Second, we discuss some open problems of modelling MPOs relevant to our approach. Third, we conclude with an outlook.

7.1 *Evaluation*

7.1.1 Descriptive statistics. Common descriptive statistics applied to all available data of a single individual or of a group are unsuitable to investigate motion. Collapsing the data into a set of descriptive measures makes it impossible to detect inter-object relations and spatially or temporally delimited motion patterns. By contrast, the REMO analysis concept allows the data of many individuals to be investigated concurrently and thus allows detection of short- and long-term as well as inter-object relationships.

7.1.2 Database management systems. In database research, numerous approaches exist to extend common query languages to cover the special properties of spatio-temporal data, e.g. SQL/temporal, (TQuel, TSQL2) or Future Temporal Logic (TFL) (Snodgrass 1987, Snodgrass and Kucera 1995, Sistla *et al.* 1998, Abraham and Roddick 1999). Such temporal queries may even involve moving query windows (Raptopoulou *et al.* 2003). Database queries and the REMO syntactic pattern detection approach have in common that the users must have a predefined idea of what they are looking for. However, querying a database normally implies retrieving stored objects, collections of objects or their observations from a database. Potential results from a database query are single MPOs or lists of MPOs, temporal intervals or subspaces that satisfy specific spatio-temporal conditions. By contrast, REMO performs an analysis on information that is not *per se* stored in the database. The REMO approach allows not only query but interrelation of stored data and derived data to create value-added information about motion events. Motion events are intrinsically spatio-temporal entities, relating the individual histories of single MPOs with the collective history of a group (cf. the trend-setter pattern).

Furthermore, the REMO approach allows patterns to be constructed on a higher level of abstraction, describing only the structural hull of an event but not its specific instantiation. For instance, $P = I([\#]\{3\})$ describes only the abstract form of a motion behaviour, matching three MPOs heading in the same direction, irrespective of whether this is north, east, west, or south. Since REMO works on derivatives of the actual stored data, it can perform an analysis on different granularities. The granularity can easily be changed by the potential user for every analysis session. Varying the granularity of the pattern-matching process can even be an important analysis task.

7.1.3 Dynamic cartography and exploratory spatial data analysis (ESDA). Another set of analysis tools for tracking data focuses on the visualization of spatio-temporal data and on analysis based on interactivity (e.g. Andrienko and Andrienko 1999, MacEachren *et al.* 1999, Edsall *et al.* 2000). ESRI Inc., for instance, offers a tool to visualize and analyse tracking data, the ArcGIS Tracking Analyst extension. It features various symbology options and a sophisticated playback manager. Its exploratory power lies, however, in the functionality to define events and to visualize where and when they occur. Thus, an analysis is performed in an exploratory way, depending predominantly on the user's knowledge of the data and their sensitivity to conspicuous features. Most of the ESDA approaches share the *data projection and reduction* step with the REMO analysis concept. However, while dynamic cartography and ESDA depend on an alert user to find eye-catching patterns or trends viewing the data from varying perspectives, the REMO analysis concept provides a quantitative approach. Instead of qualitative exploration, it offers the formalization of expected patterns and their detection in a quantitative and automated and, above all, in an objective and repeatable way. In short, it adopts the syntactic pattern detection approach.

7.1.4 KDD and data mining. KDD and its component data mining in spatio-temporal data are mainly focused on cluster detection in changing point distributions. Besides the seminal work of Openshaw (Openshaw 1994, Openshaw *et al.* 1999) on long-term disease data, further research was also carried out. For instance, Sadahiro worked on various urban point patterns (Sadahiro 2002). Whereas Openshaw's approaches follow the 'search everywhere for the unusual' philosophy, Sadahiro tracks the local maxima of density surfaces over time to reveal the displacement of disease clusters. However, there is a fundamental difference between these approaches and the REMO analysis concept: In the former case, the points represent point occurrences of, for instance, disease cases, that is point distributions without trackable individuals. In the latter case, the points represent successive fixes of trackable individuals. The latter information is of potentially higher information content and offers completely different analysis approaches, e.g. the establishment of spatio-temporal inter-object relationships shown in the REMO analysis concept.

With the REMO analysis concept, we propose a comprehensive procedure for GKD in MPO data. Our approach incorporates the whole range of KDD functions introduced by Fayyad *et al.* (1996) and adapted for geographic knowledge discovery by Miller and Han (2001): data selection, data pre-processing, data enrichment, data reduction and projection, data mining and visualization, interpretation and reporting. The REMO analysis process involves first the selection and integration of lifeline data (cf. data selection, pre-processing), second the derivation of motion properties using the detached attribute functions (cf. data enrichment), third the construction of the REMOMatrix (cf. data reduction and projection), fourth the application of the REMO data-mining algorithms (cf. data mining), and fifth the use of the visualization tools to facilitate interpretation (cf. visualization, interpretation and reporting).

Patterns are non-random properties and relationships that are valid, novel, useful and ultimately understandable (Fayyad *et al.* 1996). *Valid* means that the pattern has to be general enough to apply to new data. The REMO patterns have been tested with such diverse data as GPS-tracked animals, football players on a pitch and moving data points in an abstract ideological space. A *novel* pattern is non-trivial and unexpected. A trend-setter anticipating 12 time steps before 45 followers join a motion feature is without doubt both non-trivial and unexpected. To be *useful*, a pattern has to lead to some benefit for the user. Even though a list of REMO pattern matches does not explain the complex behaviour of wild animals, the sometimes unpredictable paths of football players, or the complex processes of Swiss society, it provides useful initial insights and gives hints for further scientific investigations. Finally, a pattern should be *ultimately understandable*, that is, simple and interpretable by humans. Using obvious and intuitive patterns, a simple analysis matrix and an extension of an established formalism, the REMO analysis concept is easy to understand for its potential users.

7.2 *Open problems*

The integration of the MPO MODELLING DOMAIN and the ANALYSIS DOMAIN revealed a set of generic problems of describing the motion of tracked individuals. Since similar problems arise in any implementation of MPO models, we discuss them here in detail.

7.2.1 Uncertain and missing fixes. Dealing with real tracking data, one is often faced with uncertain or missing fixes. For the detection of relative motion patterns, intervals with uncertain or missing MPO tracks can be fatal. Within the REMO approach, the use of detached attribute functions to construct a derivative analysis matrix at arbitrary granularities allows us to circumvent this problem. Nevertheless, real geospatial lifeline data will always be incomplete and error-prone. Thus, dealing with imperfect lifeline data remains an open research problem (Pfoser and Jensen 1999, Wentz *et al.* 2003).

7.2.2 Interpolation issues. Just as in the spatial case (see O'Sullivan and Unwin 2003), radius-limited or nearest-neighbour interpolation of motion attributes on lifeline data have two points in common. First, the size of the moving window is arbitrary. Even though users will select the interval of the moving window according to the characteristics of the data, different interval lengths may change the results of the data-mining process. Second, the wider the moving window is chosen, the smoother the lifeline description becomes. In the case of tracking data with a very high sampling rate, this may be desired to eliminate the effects of inaccurate fixes. In many other cases, excessive smoothing undesirably blurs the crucial turns of lifelines.

7.2.3 Fix sampling rate vs. analysis granularity. As introduced in section 3.1, oversampling lifelines with sparse fixes may create artefacts, e.g. false constancy patterns filling the gaps between temporally distant fixes. To avoid semantic mismatches, the tracking data should have the same sampling rate as the granularity used in the analysis task (Cheylan 2001, Hornsby 2001). Thus, the temporal granularity of the sequences should be chosen at the granularity given by the fix sampling rate. On the other hand, the Nyquist–Shannon sampling theorem (Nyquist 1928, Shannon 1949) applies if every single fix must be sampled. Following the theorem, the sampling frequency for computing motion attributes must be greater than twice the shortest interval between two fixes. Since the detached attribute functions use interpolation, the Nyquist–Shannon theorem, however, cannot be applied in the REMO approach. Nevertheless, integrating the granularity problem and the sampling theorem a rule of thumb can be stated. The motion attribute granularity should be similar to the sampling rate of the original observation data (Turchin 1998). In short, the sampling rate of the input data limits the analysis.

7.2.4 Aggregation. Another crucial issue arises similar to the modifiable areal unit problem (MAUP) (Openshaw 1984). Describing lifelines, we consider different temporal aggregations instead of spatial aggregations as with the classical MAUP. The potentially arbitrary aggregation comes with the mapping of the irregular fixes on to the regular REMO matrix, determined by a granularity and a starting time. If the temporal units were specified differently, we might observe very different patterns and relationships. The effects of different aggregation schemes on lifeline data will therefore have to be studied in detail.

7.2.5 Classification. Yet another granularity effect shows the classification of the motion attributes. The number of matched patterns is highly dependent on the attribute granularity of the pattern matching process. For example, the classification of motion azimuth into only the two classes *east* and *west* would reveal a lot of presumably meaningless constancy patterns. By contrast, every constancy pattern found with 360 azimuth classes would tend to be highly meaningful. Linear, standard deviation or quantile approaches all produce different classification results and thus different patterns and relationships.

7.3 *Outlook*

In future work, we intend to test the REMO approach with an extended set of real case studies and synthetic data. In addition to the football players and sociopolitical data described in the case study section, we will include data of GPS tracked ungulates and sharks. To create synthetic data, we are using constrained random walk (CRW) models. CRW models make it possible to simulate geospatial lifelines of arbitrary temporal and spatial granularity for stochastic numerical experiments, i.e. Monte Carlo (MC) simulations.

The aim of the MC experiments is to establish a more objective basis for deciding whether or not the detected REMO patterns are random and, if the patterns are not random, to quantify how unusual they are. Therefore, we will first generate a set of n sample data sets and subsequently run the REMO data-mining algorithms on the generated lifelines, systematically varying the settings of the pattern-matching process. We therefore can evaluate the meaningfulness and interestingness (Silberschatz and Tuzhilin 1995, 1996) of our patterns and discover potential granularity effects.

We will furthermore continue and intensify the investigation of further and more complex motion parameters such as speed, change of speed, and motion sinuosity. Additionally, we will address in detail the problem of interpolating and aggregating lifeline data. We will therefore integrate additional interpolation and aggregation techniques into the detached attribute functions and investigate their performance with our case-study data.

8. Conclusions

The main contribution of this paper is the development of a generic approach for geographic knowledge discovery (GKD) in geospatial lifeline data. The approach builds on the integration of the following crucial steps of knowledge discovery in databases (KDD) (Fayyad *et al.* 1996, Miller and Han 2001):

- *Data reduction and projection:* The transformation of the lifeline data to an analysis matrix featuring motion attributes (i.e. speed, change of speed, motion azimuth) allows the comparison of the object's motion across objects and over time.
- *Exploratory analysis and model selection:* The REMO formalism allows the description of patterns of relative motion in geospatial lifeline data.
- *Data mining:* Pattern-detection algorithms facilitate the automatic search for relative motion patterns within large data sets.
- *Visualization:* Interactive linking of the mined patterns with the object's motion in a MPO viewer allows the interpretation of the mined patterns in order to suggest avenues for further research.

To demonstrate the generic nature of the approach, we used football players tracked on a pitch and data points moving in an abstract ideological space, namely displacement of Swiss districts in ideological space over 20 years. We demonstrated that our proposed methodology is able to extract a set of valid, novel, useful and understandable motion patterns from these datasets.

We showed that REMO GKD reveals many more motion patterns than are easily identified through visual inspection and can do so in a more formal and hence repeatable way. Thus, the presented methodology may increase our knowledge about various processes in dynamic geospatial data. More generally speaking, GKD is an interesting and promising analysis technique to respond to the current increase of spatio-temporal data and the developing need for their analysis.

Acknowledgements
This work would not have been possible without the support of the data providers. Above all, the authors would like to thank Sachiko Iwase, Hideo Saito, and Shinji Ozawa of Keio University, Japan, for providing the splendid football tracking data. We thank Michael Hermann and Heiri Leuthold for providing the Swiss district data and for help in interpreting the data-mining output. Thanks to Ross Purves for his proof-reading and his helpful comments on this research. Furthermore, we are indebted to three anonymous reviewers.

References

Abraham, T. and Roddick, J.F., 1999, Survey of spatio-temporal databases. *GeoInformatica*, **3**, pp. 61–99.

Andrienko, N.V. and Andrienko, G.L., 1999, Interactive maps for visual data exploration. *International Journal of Geographical Information Science*, **13**, pp. 355–374.

ARONOFF, S., 1989, *Geographic Information Systems: A Management Perspective* (Ottawa, Canada: WDL Publications).

BATTY, M., DESYLLAS, J. and DUXBURY, E., 2003, The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades. *International Journal of Geographical Information Science*, **17**, pp. 673–697.

BÉDARD, Y., GOSSELIN, P., RIVESTA, S., PROULX, M.-J., NADEAU, M., LEBEL, G. and GAGNON, M.-F., 2003, Integrating GIS components with knowledge discovery technology for environmental health decision support. *International Journal of Medical Informatics*, **70**, pp. 79–94.

BETTINI, C., JAJODIA, S. and WANG, S., 2000, *Time Granularities in Databases, Data Mining, and Temporal Reasoning*, 1st edition (Berlin: Springer).

CHENG, R., KALASHNIKOW, D.V. and PRABHAKAR, S., 2004, Querying imprecise data in moving object environments. *IEEE Transactions on Knowledge and Data Engineering*, **16**, pp. 1112–1127.

CHEYLAN, J.-P., 2001, Time and spatial database, a conceptual application framework. In *GISDATA8, Life and Motion of Socio-Economic Units*, A. Frank, J. Raper and J. Cheylan (Eds), pp. 35–48 (London: Taylor & Francis).

EDSALL, R.M., HARROWER, M. and MENNIS, J., 2000, Tools for visualizing properties of spatial and temporal periodicity in geographic data. *Computers and Geosciences*, **26**, pp. 109–118.

ERWIG, M., GÜTING, R.H., SCHNEIDER, M. and VAZIRGIANNIS, M., 1999, Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica*, **3**, pp. 269–296.

ERWIG, M. and SCHNEIDER, M., 2002, Spatio-temporal predicates. *IEEE Transactions on Knowledge and Data Engineering*, **14**, pp. 881–901.

ERWIG, M. and SCHNEIDER, M., 2003, A visual language for the evolution of spatial relationships and its translation into a spatio-temporal calculus. *Journal of Visual Languages and Computing*, **14**, pp. 181–211.

FAYYAD, U., PIATETSKY-SHAPIRO, G. and SMYTH, P., 1996, From data mining to knowledge discovery in databases. *AI Magazine*, **17**, pp. 37–54.

FRANK, A.U., 2001, Socio-economic units: Their life and motion. In *Life and Motion of Socio-Economic Units*, Vol. 8 of *GISDATA*, 1st edition, Chap. 2, A.U. Frank, J. Raper and J.-P. Cheylan (Eds), pp. 21–34 (London: Taylor & Francis).

FRANK, A.U., 2003, Ontology for spatio-temporal databases. In *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, Vol. 2520 of *Lecture Notes in Computer Science*, Chap. 2, M. Koubarakis, T. Sellis, A.U. Frank, S. Grumbach, R.H. Gueting, C.S. Jensen, N. Lorentzos, Y. Manolopoulos, E. Nardelli, B. Pernici, H.-J. Schek, M. Scholl, B. Theodoulidis and N. Tryfona (Eds), pp. 9–77 (Berlin: Springer).

FRIEDL, J.E.F., 2002, *Mastering Regular Expression*, 2nd edition (Beijing: O'Reilly).

GANSKOPP, D., 2001, Manipulating cattle distribution with salt and water in large arid-land pastures: a GPS/GIS assessment. *Applied Animal Behaviour Science*, **73**, pp. 251–262.

GOLLEDGE, R.G., 2002, The nature of geographic knowledge. *Annals of the Association of American Geographers*, **92**, pp. 1–14.

GOODRICH, M.T. and TAMASSIA, R., 1998, *Data Structures and Algorithms in Java* (New York: Wiley).

GRUMBACH, S., KOUBARAKIS, M., RIGAUX, P., SCHOLL, M. and SKIADOPOULOS, S., 2003, Spatio-temporal models and languages: An approach based on constraints. In *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, Vol. 2520 of *Lecture Notes in Computer Science*, Chap. 5, M. Koubarakis, T. Sellis, A.U. Frank, S. Grumbach, R.H. Gueting, C.S. Jensen, N. Lorentzos, Y. Manolopoulos, E. Nardelli, B. Pernici, H.-J. Schek, M. Scholl, B. Theodoulidis and N. Tryfona (Eds), pp. 177–201 (Berlin: Springer).

Grumbach, S., Rigaux, P. and Segoufin, L., 2001, Spatio-temporal data handling with constraints. *GeoInformatica*, **5**, pp. 95–115.

Gueting, R.H., Boehlen, M.H., Erwig, M., Jensen, C.S., Lorentzos, N., Nardelli, E., Schneider, M. and Viqueira, J.R.R., 2003, Spatio-temporal models and languages: An approach based on data types. In *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, Vol. 2520 of *Lecture Notes in Computer Science*, Chap. 4, M. Koubarakis, T. Sellis, A.U. Frank, S. Grumbach, R.H. Gueting, C.S. Jensen, N. Lorentzos, Y. Manolopoulos, E. Nardelli, B. Pernici, H.-J. Schek, M. Scholl, B. Theodoulidis and N. Tryfona (Eds), pp. 117–176 (Berlin: Springer).

Hermann, M. and Leuthold, H., 2001, Weltanschauungen und ihre soziale Basis im Spiegel der Volksabstimmungen. *Swiss Political Science Review*, **7**, pp. 39–63.

Hermann, M. and Leuthold, H., 2003, *Atlas der politischen Landschaften. Ein weltanschauliches Portrait der Schweiz* (Zurich: vdf-Verlag).

Hornsby, K., 2001, Temporal zooming. *Transactions in GIS*, **5**, pp. 255–272.

Hornsby, K. and Egenhofer, M.J., 2002, Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, **36**, pp. 177–194.

Huisman, O. and Forer, P., 1998, Computational agents and urban life spaces: a preliminary realisation of the time-geography of student lifestyles. In *Proceedings of the 3rd International Conference of GeoComputation*, University of Bristol, 17–19th September, Bristol, UK, "GeoComputation CD-ROM", ISBN 0-9533477-0-2, http://www.geocomputation.org/1998/68/gc_68a.htm

Iwase, S. and Saito, H., 2002, Tracking soccer player using multiple views. In *IAPR Workshop on Machine Vision Applications* (MVA Proceedings), pp. 102–105.

Iwase, S. and Saito, H., 2003, Tracking soccer players based on homography among multiple views. In *Visual Communications and Image Processing* (Proceedings of SPIE), T. Ebrahimi and T. Sikora (Eds), pp. 288–292.

Jain, A.K., Duin, R.P.W. and Mao, J., 2000, Statistical pattern recognition: A review. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, **22**, pp. 4–37.

Knuth, D.E., Morris, J.H. and Pratt, V.R., 1977, Fast pattern matching in strings. *SIAM Journal on Computing*, **6**, pp. 323–350.

Koubarakis, M., Theodoridis, Y. and Sellis, T., 2003, Spatio-temporal databases in the years ahead. In *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, Vol. 2520 of *Lecture Notes in Computer Science*, Chap. 9, M. Koubarakis, T. Sellis, A.U. Frank, S. Grumbach, R.H. Gueting, C.S. Jensen, N. Lorentzos, Y. Manolopoulos, E. Nardelli, B. Pernici, H.-J. Schek, M. Scholl, B. Theodoulidis and N. Tryfona (Eds), pp. 345–347 (Berlin: Springer).

Kraak, M.-J. and Koussoulakou, A., 2004, A visualization environment for the space-time-cube. In *Developments in Spatial Data Handling, Proceedings of the 11th International Symposium on Spatial Data Handling, 23rd–25th August, Leicester, UK*, P.F. Fisher (Ed.), pp. 189–200 (Berlin: Springer).

Laube, P. and Imfeld, S., 2002, Analyzing relative motion within groups of trackable moving point objects. In *Geographic Information Science, Second International Conference, GIScience 2002, Boulder, CO, USA, September 2002*, Vol. 2478 of *Lecture Notes in Computer Science*, M.J. Egenhofer and D.M. Mark (Eds), pp. 132–144 (Berlin: Springer).

Laube, P., van Kreveld, M. and Imfeld, S., 2004, Finding REMO—detecting relative motion patterns in geospatial lifelines. In *Developments in Spatial Data Handling, Proceedings of the 11th International Symposium on Spatial Data Handling*, P.F. Fisher (Ed.), pp. 201–214 (Berlin: Springer).

MacEachren, A.M., Wachowicz, M., Edsall, R. and Haug, D., 1999, Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science*, **13**, pp. 311–334.

MARK, D.M., 1998, Geospatial lifelines. In *Integrating Spatial and Temporal Databases*, 98471 (Dagstuhl Seminars).

MARK, D.M., 2003, Geographic information science: Defining the field. In *Foundations of Geographic Information Science*, Chap. 1, M. Duckham, M.F. Goodchild and M.F. Worboys (Eds), pp. 3–18 (London: Taylor & Francis).

MILLER, H.J., 1991, Modelling accessibility using space–time prism concepts within geographical information systems. *International Journal of Geographical Information Systems*, **5**, pp. 287–301.

MILLER, H.J., 2003, What about people in geographic information science? *Computers, Environment and Urban Systems*, **27**, pp. 447–453.

MILLER, H.J. and HAN, J., 2001, Geographic data mining and knowledge discovery: An overview. In *Geographic Data Mining and Knowledge Discovery*, H.J. Miller and J. Han (Eds), pp. 3–32 (London: Taylor & Francis).

MILLER, H.J. and WU, Y.-H., 2000, GIS software for measuring space–time accessibility in transportation planning and analysis. *GeoInformatica*, **4**, pp. 141–159.

MOUNTAIN, D. and RAPER, J., 2001, Modelling human spatio-temporal behaviour: A challenge for location-based services. In *Proceedings of the 6th International Conference of GeoComputation*, University of Queensland, Brisbane, Australia, 24–26th September 2001, "GeoComputation CD-ROM", ISBN 1864995637, http://www.geocomputation.org/2001/papers/mountain.pdf

NYQUIST, N., 1928, Certain topics in telegraph transmission theory. *Trans. AIEE*, **47**, pp. 617–644.

OPENSHAW, S., 1984, *The Modifiable Areal Unit Problem* (Norwich, UK: Geo Books).

OPENSHAW, S., 1994, Two exploratory space-time-attribute pattern analysers relevant to GIS. In *GIS and Spatial Analysis*, Chap. 5, S. Fotheringham and P. Gogerson (Eds), pp. 83–104 (London: Taylor & Francis).

OPENSHAW, S., TURTON, I. and MAGILL J., 1999, Using geographic analysis machine to analyze limiting long-term illness cencus data. *Geographical and Environmental Modelling*, **3**, pp. 83–99.

O'SULLIVAN D. and UNWIN, D.J., 2003, *Geographic Information Analysis*, 1st edition (Hoboken, NJ: Wiley).

PFOSER, D. and JENSEN, C.S., 1999, Capturing the uncertainty of moving-object representations. In *Advances in Spatial Databases, 6th International Symposium, SSD'99, Hong Kong, China, July 1999. LNCS 1651*, R.H. Gueting, D. Papadias and F. Lochowsky (Eds), pp. 111–131 (Berlin: Springer).

RAPER, J. and LIVINGSTONE, D., 1995, Development of a geomorphological spatial model using object-oriented design. *International Journal of Geographical Information Systems*, **9**, pp. 359–383.

RAPTOPOULOU, K., PAPADOPOULOS, A.N. and MANOLOPOULOS, Y., 2003, Fast nearest-neighbor query processing in moving-point databases. *GeoInformatica*, **7**, pp. 113–137.

RIVEST, S., BÉDARD, Y. and MARCHAND, P., 2001, Towards a better support for spatial decision making: Defining the characteristics of spatial on-line analytical processing (SOLAP). *Geomatica*, **55**, pp. 539–555.

RODDICK, J.F., HORNSBY, K. and SPILIOPOULOU, M., 2001, An updated bibliography of temporal, spatial, and spatio-temporal data mining research. In *Temporal, Spatial and Spatio-Temporal Data Mining, TSDM 2000*, Vol. 2007 of *Lecture Notes in Artificial Intelligence*, J.F. Roddick and K. Hornsby (Eds), pp. 147–163 (Berlin: Springer).

ROOT, R.B. and KAREIVA, P.M., 1984, The search for resources by cabbage butterflies (*Pieris rapae*): ecological consequneces and adaptive significance of markovian movement in a patchy environment. *Ecology*, **65**, pp. 147–165.

SADAHIRO, Y., 2002, A graphical method for exploring spatiotemporal point distributions. *Cartography and Geographic Information Science*, **29**, pp. 67–84.

SHANNON, C., 1949, Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers*, **37**, pp. 10–21.

SIBBALD, A.M., HOOPER, R.J., GORDON, I.J. and CUMMING, S., 2001, Using GPS to study the effect of human disturbance on the behaviour of the red deer stags on a highland estate in Scotland. In *Tracking Animals with GPS*, A.M. Sibbald and I.J. Gordon (Eds), pp. 39–43 (Aberdeen, UK: Macaulay Institute).

SILBERSCHATZ, A. and TUZHILIN, A., 1995, On subjective measures of interestingness in knowledge discovery. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, Montreal*, pp. 275–281.

SILBERSCHATZ, A. and TUZHILIN, A., 1996, What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, **8**, pp. 970–974.

SISTLA, A.P., WOLFSON, O., CHAMBERLAIN, S. and DAO, S., 1997, Moving objects databases: Issues and solutions. In *Proceedings of the Thirteenth International Conference on Data Engineering (ICDE13)*, April 1997 (Birmingham, UK).

SISTLA, A.P., WOLFSON, O., CHAMBERLAIN, S. and DAO, S., 1998, Querying the uncertain position of moving objects. In *Temporal Databases—Research and Practice*, Vol. 1399 of *Lecture Notes in Computer Science*, O. Etzion, S. Jajodia and S. Sripada (Eds), pp. 310–337 (Berlin: Springer).

SMYTH, C.S., 2001, Mining mobile trajectories. In *Geographic Data Mining and Knowledge Discovery*, H.J. Miller and J. Han (Eds), pp. 337–361 (London: Taylor & Francis).

SNODGRASS, R.T., 1987, The temporal query language TQuel. *ACM Transactions on Database Systems*, **12**, pp. 247–298.

SNODGRASS, R.T. and KUCERA, H., 1995, Rationale for a temporal extension of SQL. In *The TSQL2 Temporal Query Language*, pp. 3–18 (Dordrecht, The Netherlands: Kluwer).

TRAJCEVSKI, G., WOLFSON, O., HINRICHS, K. and CHAMBERLAIN, S., 2004, Managing uncertainty in moving objects databases. *ACM Transactions on Database Systems (TODS)*, **29**, pp. 463–507.

TURCHIN, P., 1991, Translating foraging movements in heterogeneous environments into the spatial distribution of foragers. *Ecology*, **72**, pp. 1253–1266.

TURCHIN, P., 1998, *Quantitative Analysis of Movement: Measuring and Modelling Population Redistribution in Animals and Plants* (Sunderland, MA: Sinauer).

WALL, L., CHRISTIANSEN, T. and SCHWARTZ, R.L., 1996, *Programming Perl*, 2nd edition (Sebastopol, CA: O'Reilly).

WENTZ, E.A., CAMPELL, A.F. and HOUSTON, R., 2003, A comparision of two methods to create tracks of moving objects: linear weighted distance and constrained random walk. *International Journal of Geographical Information Science*, **17**, pp. 623–645.

WOLFSON, O., XU, B., CHAMBERLAIN, S. and JIANG, L., 1998, Moving objects databases: Issues and solutions. In *10th International Conference on Scientific and Statistical Database Management, Proceedings*, Capri, Italy, July 1–3, 1998 (IEEE Computer Society), M. Rafanelli and M. Jarke (Eds), pp. 111–131, http://csdl.computer.org/comp/proceedings/ssdbm/1998/8575/00/85750111abs.htm

Finding REMO - Detecting Relative Motion Patterns in Geospatial Lifelines

Patrick Laube[1], Marc van Kreveld[2], and Stephan Imfeld[1]

Department of Geography, University of Zurich, Winterthurerstrasse 190, CH–8057 Zurich, Switzerland, {plaube,imfeld}@geo.unizh.ch[1]
Department of Computer Science, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, The Netherlands, marc@cs.uu.nl[2]

Abstract

Technological advances in position aware devices increase the availability of tracking data of everyday objects such as animals, vehicles, people or football players. We propose a geographic data mining approach to detect generic aggregation patterns such as flocking behaviour and convergence in geospatial lifeline data. Our approach considers the object's motion properties in an analytical space as well as spatial constraints of the object's lifelines in geographic space. We discuss the geometric properties of the formalised patterns with respect to their efficient computation.
Keywords: Convergence, cluster detection, motion, moving point objects, pattern matching, proximity

1 Introduction

Moving Point Objects (MPOs) are a frequent representation for a wide and diverse range of phenomena: for example animals in habitat and migration studies (e.g. Ganskopp 2001; Sibbald et al. 2001), vehicles in fleet management (e.g. Miller and Wu 2000), agents simulating people for modelling crowd behaviour (e.g. Batty et al. 2003) and even tracked soccer players on a football pitch (e.g. Iwase and Saito 2002). All those MPOs share motions that can be represented as *geospatial lifelines*: a series of observations consisting of a triple of *id*, *location* and *time* (Hornsby and Egenhofer 2002).

Gathering tracking data of individuals became much easier because of substantial technological advances in position aware devices such as GPS receivers, navigation systems and mobile phones. The increasing number of such devices will lead to a wealth of data on space-time trajectories documenting the space-time behaviour of animals, vehicles and people for off-line analysis. These collections of geospatial lifelines present a rich environment to analyse individual behaviour. (Geographic) data mining may detect patterns and rules to gather basic knowledge of dynamic processes or to design location based services (LBS) to simplify individual mobility (Mountain and Raper 2001; Smyth 2001; Miller 2003).

Knowledge discovery in databases (KDD) and data mining are responses to the huge data volumes in operational and scientific databases. Where traditional analytical and query techniques fail, data mining attempts to distill data into information and KDD turns information into knowledge about the monitored world. The central belief in KDD is that information is hidden in very large databases in the form of interesting patterns (Miller and Han 2001). This statement is true for the spatio-temporal analysis of geospatial lifelines and thus is a key motivator for the presented research. Motion patterns help to answer the following type of questions.

- Can we identify an alpha animal in the tracking data of GPS-collared wolves?
- How can we quantify evidence of 'swarm intelligence' in gigabytes of log-files from agent-based models?
- How can we identify which football team played the more catching lines of defense in the lifelines of 22 players sampled at seconds?

The long tradition of data mining in the spatio-temporal domain is well documented (for an overview see Roddick et al. (2001)). The Geographic Information Science (GISc) community has recognized the potential of Geographic Information Systems (GIS) to 'capture, represent, analyse and explore spatio-temporal data, potentially leading to unexpected new knowledge about interactions between people, technologies and urban infrastructures (Miller 2003). Unfortunately, most commercial GIS are based on a static place-based perspective and are still notoriously weak in providing tools for handling the temporal dimensions of geographic information (Mark 2003). Miller postulates expanding GIS from the place-based perspective to encompass a people-based perspective. He identifies the development of a formal representation theory for dynamic spatial objects and of new spatio-temporal data mining and exploratory visualization techniques as key research issues for GISc.

In this paper work is presented which extends a concept developed to analyse relative motion patterns for groups of MPOs (Laube and Imfeld 2002) to also analyse the object's absolute locations. The work allows the identification of generic formalised motion patterns in tracking data and the extraction of instances of these formalised patterns. The significance of these patterns is discussed.

2 Aggregation in Space and Time

Following Waldo Tobler's first law of geography, near things are more related than distant things (Tobler 1970). Tolber's law is often referred to as being the core of spatial autocorrelation (Miller 2004). Nearness as a concept can be extended to include both space and time. Thus analysing geospatial lifelines we are interested in objects near in space-time. Objects that are near at certain times might be related. Although correlation is not causality, it provides evidence of causality that can (and should) be assessed in the light of theory and/or other evidence. Since this paper focuses on the formal and geometrical definition and the algorithmic detection of motion patterns we use geometric *proximity* in euclidian space to avoid the vague term nearness.

To analyse geospatial lifelines this could mean that MPOs moving within a certain range influence each other. E.g. an alpha wolf leads its pack by being seen or heard, thus all wolves have to be located within the range of vision or earshot respectively. Analysing geospatial lifelines we are interested in first identifying motion patterns of individuals moving in proximity. Second we want to know how, when and where sets of MPOs aggregate, converge and build clusters respectively.

Investigating aggregation of point data in space and time is not new. Most approaches focus on detecting localized clustering at certain time slices (e.g. Openshaw 1994; Openshaw et al. 1999). This concept of spatial clusters is static, rooted in the time sliced static map representation of the world. With a true spatio-temporal view of the world aggregation must be perceived as the momentary process convergence and the final static cluster as its possible result. The opposite of convergence, divergence, is equally interesting. Its possible result, some form of dispersal, is much less obvious and thus much harder to perceive and describe.

A cluster is not the compulsory outcome of a convergence process and vice versa. A set of MPOs can very well be converging for a long time without building a cluster. The 22 players of a football match may converge during an attack without ever forming a detectable cluster on the pitch. In reverse, MPOs moving around in a circle may build a wonderful cluster but never be converging. In addition the process of convergence and the final cluster are in many cases sequential. Consider the lifelines of a swarm of bees. At sunset the bees move back to the hive from the surrounding meadows, showing a strong convergence pattern without building a spatial cluster. In the hive the bees wiggle around in a very dense cluster, but do not converge anymore. In short, even though convergence and clustering are often spatially and/or temporally tied up, there need not be a detectable relation in an individual data frame under investigation.

3 The Basic REMO–Analysis Concept

The basic idea of the analysis concept is to compare the motion attributes of point objects over space and time, and thus to *relate* one object's motion to the motion of all others (Laube and Imfeld 2002). Suitable geospatial lifeline data consist of a set of MPOs each featuring a list of fixes. The REMO concept (RElative MOtion) is based on two key features: First, a transformation of the lifeline data to a REMO matrix featuring motion attributes (i.e. speed, change of speed or motion azimuth). Second, matching of formalized patterns on the matrix (Fig. 1).

Two simple examples illustrate the above definitions: Let the geospatial lifelines in Fig. 1a be the tracks of four GPS-collared deer. The deer O_1 moving with a constant motion azimuth of 45° during an interval t_2 to t_5, i.e. four discrete time steps of length ∂t, is showing *constance*. In contrast, four deer performing a motion azimuth of 45° at the same time t_4 show *concurrence*.

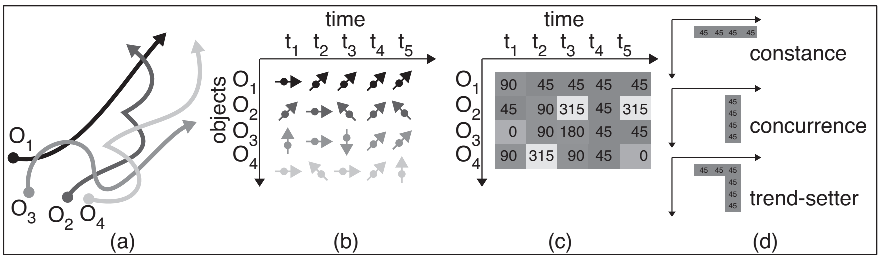


Fig. 1. The geospatial lifelines of four MPOs (a) are used to derive in regular intervals the motion azimuth (b). In the REMO analysis matrix (c) generic motion patterns are matched (d).

The REMO concept allows construction of a wide variety of motion patterns. See the following three basic examples:

- *Constance:* Sequence of equal motion attributes for *r* consecutive time steps (e.g. deer O_1 with motion azimuth 45° from t_2 to t_5).
- *Concurrence:* Incident of *n* MPOs showing the same motion attributes value at time *t* (e.g. deer O_1, O_2, O_3, and O_4 with motion azimuth 45° at t_4)
- *Trend-setter:* One trend-setting MPO anticipates the motion of *n* others. Thus, a trend-setter consists of a *constance* linked to a *concurrence* (e.g. deer O_1 anticipates at t_2 the motion azimuth 45° that is reproduced by all other MPOs at time t_4)

For simplicity we focus in the remainder of this paper on the motion attribute azimuth, even though most facets of the REMO concept are equally valid for speed or change of speed.

4 Spatially Constrained REMO patterns

The construction of the REMO-matrix is an essential reduction of the information space. However it should be noted that this step factors out the absolute locations of the fixes of the MPOs. The following two examples illustrate generic motion patterns where the absolute locations must be considered.

- Three geese all heading north-west at the same time – one over London, one over Cardiff and one over Glasgow are unlikely to be influenced by each other. In contrast, three geese moving north-west in the same gaggle are probably influenced. Thus, for flocking behaviours the spatial proximity of the MPOs has to be considered.
- Three geese all heading for Leicester at the same time – one starting over London, one over Cardiff and one over Glasgow show three different motion azimuths, not building any pattern in the REMO matrix. Thus, convergence can only be detected considering the absolute locations of the MPOs.

The basic REMO concept must be extended to detect such *spatially constrained* REMO patterns. In Section 4.1 spatial proximity is integrated and in Section 4.2 an approach is presented to detect convergence in relative motion. Section 4.3 evaluates algorithmic issues of the proposed approaches.

4.1 Relative Motion with Spatial Proximity

Many sheep moving in a similar way is not enough to define a flocking pattern. We expect additionally that all the sheep of a flock graze on the same hillside. Formalised as a generic motion pattern we expect for a flocking the MPOs to be in spatial proximity. To test the proximity of *m* MPOs building a pattern at a certain time we can compute the spatial proximity of the *m* MPO's fixes in that time frame. Following Tobler's first law, proximity among MPOs can be considered as impact ranges, or the other way around: a spatio-temporally clustered set of MPOs is evidence to suggest an interrelation among the involved MPOs.

The meaning of spatial constraint in a motion pattern is different if we consider the geospatial lifeline of a single MPO. The consecutive observations (fixes) of a single sheep building a lifeline can be tested for proximity. Thus the proximity measure constrains the spatial extent of single object's motion pattern. A constance for a GPS-collared sheep may only be meaningful if it spans a certain distance, excluding pseudo-motion caused by inaccurate fix measurements.

Different geometrical and topological measures could be used to constrain motion patterns spatially. The REMO analysis concept focuses on the following open list of geometric proximity measures.

- A first geometric constraint is the mean distance to the mean or median center (length of star plot).
- Another approach to indicate the spatial proximity of points uses the Delaunay diagram, applied for cluster detection in 2-D point sets (e.g. Estivill-Castro and Lee 2002) or for the visualisation of habitat-use intensity of animals (e.g. Casaer et al. 1999). According to the cluster detection approach two points belong to the same cluster, if they are connected by a small enough Delaunay edge. Thus, adapted to the REMO concept a second distance proximity measure is to limit the average length of the Delauney edges of a point group forming a REMO pattern.
- Proximity measures can have the form of bounding boxes, circles or ellipses (Fig. 2). The simplest way of indicating an impact range would be to specify a maximal bounding box that enclosed all fixes relevant to the pattern. Circular criteria can require enclosing all relevant fixes within radius *r* or include the constraint to be spanned around the mean or median center of the fixes. Ellipses are used to rule the directional elongation of the point cluster (major axis *a*, minor axis *b*).
- Another areal proximity measure for a set of fixes is the indication of a maximal border length of the convex hull.

Using these spatial constraints the list of basic motion patterns introduced in Section 3 can be amended with the spatially constrained REMO patterns (Fig. 2).

- *Track:* Consists of the REMO pattern constance and the attachment of spatial constraint. Definition: *constance* + spatial constraint *S*.
- *Flock:* Consists of the REMO pattern concurrence and the attachment of a spatial constraint. Definition: *concurrence* + spatial constraint *S*.
- *Leadership:* Consists of the REMO pattern trend-setter and the attachment of a spatial constraint. For example the followers must lie within the range *(∂x, ∂y)* when they join the motion of the trend-setter. Definition: *trend-setter* + spatial constraint *S*.

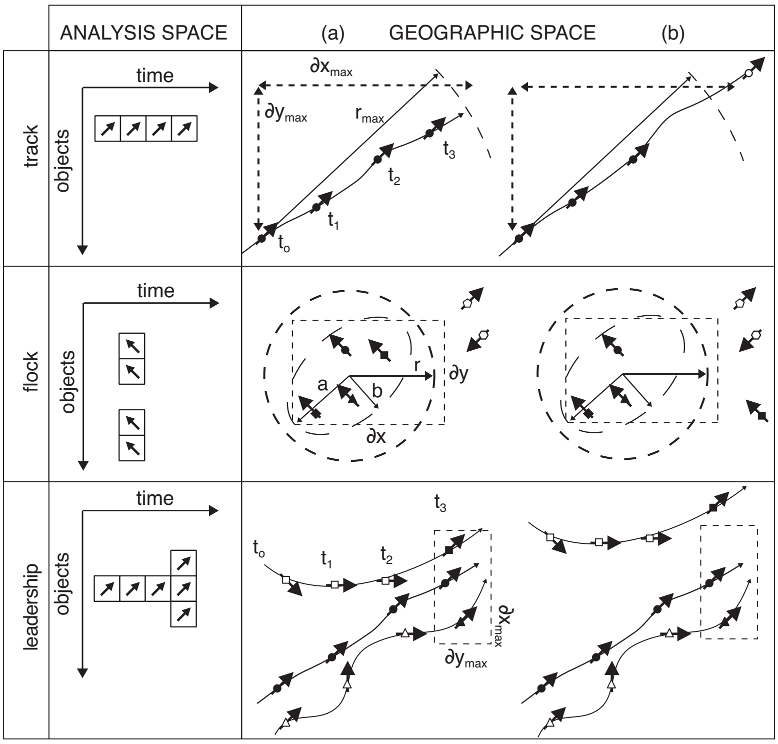


Fig. 2. The figure illustrates the constraints of the patterns track, flock and leadership in the analysis space (the REMO matrix) and in the geographic space. Fixes matched in the analysis space are represented as solid forms, fixes not matched as empty forms. Some possible spatial constraints are represented as ranges with dashed lines. Whereas in the situations (a) the spatial constraints for the absolute positions of the fixes are fulfilled they are not in the situations (b): For track the last fix lies beyond the range, for flock and leadership the quadratic object lies outside the range.

4.2 Convergence

At the same time self-evident and fascinating are groups of MPOs aggregating and disaggregating in space and time. An example is wild animals suddenly heading in a synchronised fashion for a mating place. Wildlife biologists could be interested in the *who*, *when* and *where* of this motion pattern. Who is joining this spatio-temporal trend? Who is not? When does the process start, when does it end? Where lies the mating place, what spatial extent or form does it have? A second example comes from the analysis of crowd behaviour. Can we identify points of interest attracting people only at certain times, events of interest rather than points of interest, losing their attractiveness after a while? To answer such questions we propose the spatial REMO pattern *convergence*.

The phenomenon aggregation has a spatial and a spatio-temporal form. An example may help to illustrate the difference. Let A be a set of n antelopes. A wildlife biologist may be interested in identifying sets of antelopes heading for some location at certain time. The time would indicate the beginning of the mating season, the selected set of m MPOs the ready-to-mate individuals, and the spot might be the mating area. This is a *convergence* pattern. It is primarily spatial, that means the MPOs head for an area but may reach it at different times. On the other hand the wildlife biologist and the antelopes may share the vital interest to identify MPOs that head for some location and actually meet there at some time extrapolating their current motion. Thus, the pattern *encounter* includes considerations about speed, excluding MPOs heading for a meeting range but not arriving there at a particular time with the others.

- *Convergence:* Heading for R. Set of m MPOs at interval i with motion azimuth vectors intersecting within a range R of radius r.
- *Encounter:* Extrapolated meeting within R. Set of m MPOs at interval i with motion azimuth vectors intersecting within a range R of radius r and actually meeting within R extrapolating the current motion.

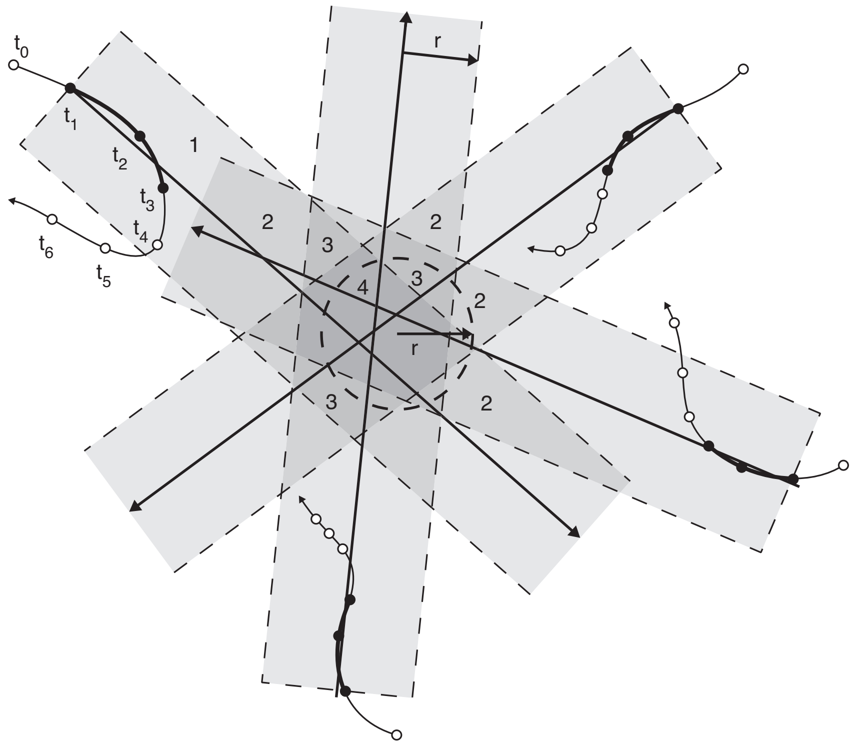


Fig. 3. Geometric detection of convergence. Let *S* be a set of 4 MPOs with 7 fixes from t_0 to t_6. The illustration shows a convergence pattern found with the parameters *4 MPOs* at the temporal interval t_1 *to* t_3. The darkest polygon denotes an area where all 4 direction vectors are passing at a distance closer than *r*. The pattern convergence is found if such a polygon exists. Please note that the MPOs do not build a cluster but nevertheless show a convergence pattern.

The convergence pattern is illustrated in Figure 3. Let *S* be a set of MPOs with *n* fixes from t_0 to t_{n-1}. For every MPO and for every interval of length *i* an azimuth vector fitting in its fixes within *i* represents the current motion. The azimuth vector can be seen as a half-line projected in the direction of motion. The convergence is matched if there is at any time a circle of radius *r* that intersects *n* directed half-lines fitted for each MPO in the fixes within *i*. For the encounter pattern whether the objects actually meet in future must additionally be tested.

The opposites of the above described patterns are termed *divergence* and *breakup*. The latter term integrates a spatial divergence pattern with the temporal constraint of a precedent meeting in a range *R*. The graphical representation of the divergence pattern is highly similar to Fig. 3. The only difference lies in the construction of the strips, heading backwards instead of forwards, relative to the direction of motion.

4.3 Algorithms and Implementation Issues

In this section we develop algorithms to detect the above introduced motion patterns and analyse their efficiency.

The basic motion patterns in the REMO concept are relatively easy to determine in linear time. The addition of positions requires more complex techniques to obtain efficient algorithms. We analyse the efficiency of pattern discovery for track, flock, leadership, convergence, and encounter in this section. We let the range be a circle of given radius R. Let n denote the number of MPOs in the data set, and t the number of time steps. The total input size is proportional to nt, so a linear time algorithm requires $O(nt)$ time. We let m denote the number of MPOs that must be involved in a pattern to make it interesting. Finally, we assume that the length of a time interval is fixed and given.

The addition of geographic position to the REMO framework requires the addition of geographic tests or the use of geometric algorithms. The *track* pattern can simply be tested by checking each basic constance pattern found for each MPO. If the constance pattern also satisfies the range condition, a track pattern is found. The test takes constant additional time per pattern, and hence the detection of track patterns takes $O(nt)$ time overall.

Efficient detection of the *flock* pattern is more challenging. We first separate the input data by equal time and equal motion direction, so that we get a set of $n' \leq n$ points with the same direction and at the same time. The $8t$ point sets in which patterns are sought have total size $O(nt)$. To discover whether a subset of size at least m of the n points lie close together, within a circle of radius R, we use higher-order Voronoi diagrams. The m-th order Voronoi diagram is the subdivision of the plane into cells, such that for any point inside a cell, some subset of m points are the closest among all the points. The number of cells is $O(m(n'-m))$ (Aurenhammer 1991), and the smallest enclosing circle of each subset of m points can be determined in $O(m)$ time (de Berg et al. 2000, Sect. 4.7). If the smallest enclosing circle has radius at most R, we have discovered a pattern. The sum of the n' values over all $8t$ point sets is $O(nt)$, so the total time needed to detect these patterns is $O(ntm^2 + nt \log n)$. This includes the time to compute the m-th order Voronoi diagram (Ramos 1999).

Leadership pattern detection can be seen as an extension of flock pattern detection. The additional condition is that one of the MPOs shows constance over the previous time steps. Leadership detection also requires $O(ntm^2+nt \log n)$ time.

For the *convergence* pattern, consider a particular time interval. The n MPOs give rise to n azimuth vectors, which we can see as directed half-lines. To test whether at least m MPOs out of n converge, we compute the arrangement formed by the thickened half-lines, which are half-strips of width $2r$. For every cell in the arrangement we determine how many thickened half-lines contribute, which can be done by traversing the arrangement once and maintaining a counter that shows in how many half-strips the current cell is. If a cell is contained in at least m half-strips, it constitutes a pattern. Computing the arrangement of n half-strips and setting the counters can be done in $O(n^2)$ time in total; the algorithm is very similar to computing levels in arrangements (de Berg et al. 2000, Chap. 8). Since we consider t different time intervals, the total running time becomes $O(n^2t)$.

The *encounter* pattern is the most complex one to compute. The reason is that extrapolated meeting times must also match, which adds a dimension to the space in which geometric algorithms are needed. We lift the problem into 3-D space, where the third dimension is time. The MPOs become half-lines that go upward from a common horizontal plane representing the beginning of the time interval; the slope of the half-lines will now be the speed. The geometric problem to be solved is finding horizontal circles of radius R that are crossed by at least m half-lines, which can be solved in $O(n^4)$ time with a simple algorithm. For all time intervals of a given length, the algorithm needs $O(n^4t)$ time.

5 Discussion

The REMO approach has been designed to analyse motion basing on geospatial lifelines. Since motion is expressed by a change in location over time the REMO patterns intrinsically span over space and time. Our approach thus overcomes the limitation of only either detecting spatial clusters on snapshots or highlighting temporal trends in attributes of spatial units. It allows pattern detection in space-time.

REMO patterns rely solely on point observations and are thus expressible for any objects that can be represented as points and leave a track in a euclidean space. Having translated the expected behaviours into REMO patterns, the detection process runs unsupervised, listing every pattern occurrence. The introduced patterns can be detected within reasonable time. Many simple patterns can be detected in close to linear time if the size of the subset m that constitutes a pattern is a constant, which is natural in many situations. The encounter pattern is quite expensive to compute, but since we focus on off-line analysis, we can still deal with data sets consisting of several hundreds of MPOs. Note that the dependency on the number of time steps is always linear for fixed length time intervals. The most promising way to obtain more efficient algorithms is by using ap*proximation algorithms*, which can save orders of magnitude by stating the computational problem slightly less firm (Bern and Eppstein 1997). In short, the REMO concept can cope with the emerging data volumes of tracking data.

Syntactic pattern recognition adopts a hierarchical perspective where complex patterns are viewed as being composed of simple primitives and grammatical rules (Jain et al. 2000). Sections 3 and 4 introduced a subset of possible pattern primitives of the REMO analysis concept. Using the primitives and a pattern description formalism almost arbitrary motion patterns can be described and detected. Due to this hierarchical design the concept easily adapts to the special requirements of various application fields. Thus, the approach is flexible and universal, suited for various lifelines such as of animals, vehicles, people, agents or even soccer players. The detection of patterns of higher complexity requires more sophisticated and flexible pattern matching algorithms than the ones available today.

The potential users of the REMO method know the phenomenon they investigate and the data describing it. Hence, in contrast to traditional data mining assuming no prior knowledge, the users come up with expectations about conceivable motion patterns and are able to assess the results of the pattern matching process. Therein lies a downside of the REMO pattern detection approach: It requires relatively sophisticated knowledge about the patterns to be searched for. For instance, the setting of an appropriate impact range for a flock pattern is highly dependent on the investigated process and thus dependent on the user. In general the parametrisation of the spatial constraints influences the number of patterns detected. Further research is needed to see whether autocalibration of pattern detection will be possible within the REMO concept.

Even though the REMO analysis concept assumes users specifying patterns they are interested in, the pattern extent can also be viewed as an analysis parameter of the data mining approach. One reason to do so is to detect scale effects lurking in different granularities of geospatial lifeline data. The number of matched patterns may be highly dependent on the spatial, temporal and attributal granularity of the pattern matching process. For example the classification of motion azimuth in only the two classes *east* and *west* reveals a lot of presumably meaningless constance patterns. In contrast, the probability of finding constance patterns with 360 azimuth classes is much smaller, or take the selection of the impact range r for the flock pattern in sheep as another example. By testing the length of the impact range *r* against the amount of matched patterns one could search for a critical maximal impact range within a flock of sheep. Future research will address numerical experiments with various data to investigate such relations.

A critical issue in detecting convergence is fitting the direction vector in a set of fixes. Only slight changes in its azimuth may have huge effects on the overlapping regions. A straightforward solution approach to this problem is to smooth the lifelines and then fitting the azimuth vector to a segment of the smoothed lifeline.

The paper illustrates the REMO concept referring to ideal geospatial lifeline data. In reality lifeline data are often imprecise and uncertain. Sudden gaps in lifelines, irregular fixing intervals or positional uncertainty of fixes require sophisticated interpolation and uncertainty considerations on the implementation side (e.g. Pfoser and Jensen 1999).

6 Conclusions

With the technology driven shift from the static map view of the world to a dynamic process in GIScience, cluster detection on snapshots is insufficient. What we need are new methods that can detect convergence processes as well as static clusters, especially if these two aspects of space-time aggregation are separated. We propose a generic, understandable and extendable approach for data mining in geospatial lifelines. Our approach integrates individuals as well as groups of MPOs. It also integrates parameters describing the motion as well as the footprints of the MPOs in space-time.

Acknowledgements

The ideas to this work prospered in the creative ambience of Dagstuhl Seminar No. 03401 on 'Computational Cartography and Spatial Modelling'. The authors would like to acknowledge invaluable input from Ross Purves, University of Zurich.

References

Aurenhammer F (1991) Voronoi diagrams: A survey of a fundamental geometric data structure. ACM Comput. Surv. 23(3):345-405

Batty M, Desyllas J, Duxbury E (2003) The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades. Int. J. Geographical Information Systems 17(7):673-697

Bern M, Eppstein D (1997) Approximation algorithms for geometric problems. In Hochbaum DS (ed) Approximation Algorithms for NP-Hard Problems, PWS Publishing Company, Boston, MA, pp 296-345

Casaer J, Hermy M, Coppin P, Verhagen R (1999) Analysing space use patterns by Thiessen polygon and triangulated irregular network interpolation: a non-parametric method for processing telemetric animal fixes. Int. J. Geographical Information Systems 13(5):499-511

de Berg M, van Kreveld M, Overmars M, Schwarzkopf O (2000) Computational Geometry - Algorithms and Applications. Springer, Berlin, 2nd edition

Estivill-Castro V, Lee I (2002) Multi-level clustering and its visualization for exploratory data analysis. GeoInformatica 6(2):123-152

Ganskopp, D. (2001) Manipulating cattle distribution with salt and water in large arid-land pastures: a GPS/GIS assessment. Applied Animal Behaviour Science 73(4):251-262

Hornsby K, Egenhofer M (2002) Modeling moving objects over multiple granularities. Annals of Mathematics and Artificial Intelligence 36(1-2):177-194

Iwase S, Saito H (2002) Tracking soccer player using multiple views. In IAPR Workshop on Machine Vision Applications, MVA Proceedings, pp 102-105

Jain A, Duin R, Mao J (2000) Statistical pattern recognition: A review. IEEE Transactions on Pattern Recognition and Machine Intelligence 22(1):4-37

Laube P, Imfeld S (2002) Analyzing relative motion within groups of trackable moving point objects. In: Egenhofer M, Mark D (eds), Geographic Information Science, Second International Conference, GIScience 2002, Boulder, CO, USA, September 2002, LNCS 2478, Springer, Berlin, pp 132-144

Mark, D (2003) Geographic information science: Defining the field. In: Duckham M, Goodchild M, Worboys M (eds), Foundations of Geographic Information Science, chap. 1, Taylor and Francis, London New York, pp 3-18

Miller H (2003) What about people in geographic information science? Computers, Environment and Urban Systems 27(5):447-453

Miller H (2004) Tobler's first law and spatial analysis. in preparation.

Miller H, Han J (2001) Geographic data mining and knowledge discovery: An overview. In: Miller H, Han J (eds) Geographic data mining and knowledge discovery, Taylor and Francis, London New York, pp 3-32

Miller H, Wu Y (2000) GIS software for measuring space-time accessibility in transportation planning and analysis. GeoInformatica 4(2):141-159

Mountain D, Raper J (2001) Modelling human spatio-temporal behaviour: A challenge for location-based services. Proceedings of GeoComputation, Brisbane, 6

Openshaw S (1994) Two exploratory space-time-attribute pattern analysers relevant to GIS. In: Fotheringham S, Gogerson P (eds) GIS and Spatial Analysis, chap. 5, Taylor and Francis, London New York, pp 83-104

Openshaw S, Turton I, MacGill J (1999) Using geographic analysis machine to analyze limiting long-term illness census data. Geographical and Environmental Modelling 3(1):83-99

Pfoser D, Jensen C (1999) Capturing the uncertainty of moving-object representations. In: Gueting R, Papadias D, Lochowsky, F (eds) Advances in Spatial Databases, 6th International Symposium, SSD'99, Hong Kong, China, July 1999. LNCS 1651, Springer, Berlin Heidelberg, pp 111-131

Ramos E (1999) On range reporting, ray shooting and k-level construction. In: Proc. 15th Annu. ACM Symp. on Computational Geometry, pp 390-399

Roddick J, Hornsby K, Spiliopoulou M (2001) An updated bibliography of temporal, spatial, and spatio-temporal data mining research. In: Roddick J, Hornsby K (eds), Temporal, spatial and spatio-temporal data mining, TSDM 2000, LNAI 2007, Springer, Berlin Heidelberg, pp 147-163

Sibbald AM, Hooper R, Gordon IJ, Cumming S (2001) Using GPS to study the effect of human disturbance on the behaviour of the red deer stags on a highland estate in Scotland. In: Sibbald A, and Gordon IJ (eds) Tracking Animals with GPS, Macaulay Institute, pp 39-43

Smyth C (2001) Mining mobile trajectories. In: Miller H, Han J (eds) Geographic data mining and knowledge discovery, Taylor and Francis, London New York, pp 337-361

Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Economic Geography 46(2):234-240

Evaluating motion pattern detection techniques in spatio-temporal data

Patrick Laube and Ross S. Purves
Department of Geography, University of Zurich
Winterthurerstrasse 190, CH-8057 Zurich
Telephone: +41-44-635 51 31
Fax: +41-44-635 68 48
{plaube,rsp}@geo.unizh.ch

submitted to *Computers, Environment and Urban Systems, 19-04-2005*

Abstract. This paper presents a method to evaluate a geographic knowledge discovery approach for exploring the motion of point objects. The goal is to provide a means of considering the significance of motion patterns, described through their *interestingness*. We use Monte-Carlo simulations of constrained random walks to generate populations of synthetic lifelines, using the statistical properties of real observational data as constraints. Pattern occurrence in the synthetic data is then compared with observational data to assess the potential interestingness of the found patterns. We use motion data from wildlife biology and spatialisation in political science for the evaluation. The results of the numerical experiments show that the interestingness of found motion patterns is largely dependant on the configuration of the pattern matching process, which includes the pattern extent, the temporal granularity, and the classification schema used for the motion attributes azimuth and speed. The results of the numerical experiments allow interestingness to be attached only to some of the patterns found, – other patterns were suggested to be not interesting. The evaluation method helps in estimating useful configurations of the pattern detection process. This work emphasises the need to further investigate the statistical aspects of the problem under study in (geographic) knowledge discovery.

Keywords. geographic knowledge discovery, motion, lifelines, pattern detection, constrained random walk, Monte Carlo experiments.

1 Introduction

Location aware devices are becoming ubiquitous and will increase our capability to collect spatio-temporal motion data by many orders of magnitude. The ubiquity of such devices is reflected by the fact that in the summer of 2004 the Japanese Telecommunications Council declared that all mobile phones introduced in Japan after 2007 should have self-locating functionality. Studies of so-called moving point objects (MPO), incorporating information about changing positions of discrete objects in time and space, have been identified as a key emerging research area in GIScience (Miller 2003). By studying MPOs through time, individual geospatial lifelines can be derived from large datasets collected, for example, from people carrying GPS-enabled phones and PDAs (e.g. Dykes & Mountain, 2003), tracked animals in field studies (e.g. Wentz et al. 2003), or even tracked football players in sports scene analysis (e.g. Iwase & Saito, 2003).

It has been recognised that not only is spatial data special, but also the handling and analysing of spatio-temporal data, and above all motion data, requires the development of new concepts (Frank, 2001; Mark, 2003). Traditional analytical methods for spatial and spatio-temporal data were developed in an era when data collection was expensive and computational power was weak (Miller & Han, 2001). Miller and Han thus reason that "traditional spatial analytical techniques cannot easily discover new and unexpected patterns, trends and relationships that can be hidden deep within very large and diverse geographic datasets" (Miller & Han, 2001, p. 3).

The integration of knowledge from the field of GIScience about space-time together with the emerging field of Knowledge Discovery in Databases opens up the possibility for Geographic Knowledge Discovery (GKD). Applications which generate large volumes of spatio-temporal data, such as high-resolution (in time and space) satellite-based systems (Griffiths & Mather, 2000), and in our case the analysis of geospatial lifelines (Hornsby & Egenhofer, 2002; Mark, 1998) face multiple challenges in the storage and exploration of high-volume spatio-temporal datasets and thus present excellent cases for the application of GKD.

It has been recognized in the knowledge discovery in databases (KDD) literature that discovery systems can generate a glut of patterns, most of which are of no interest to the user (Padmanabhan, 2004; Silberschatz and Tuzhilin, 1996). Thus, it is recommended that data mining be carried out with regard to the statistical aspects of the problem (Fayyad et al., 1996).

In this paper work is presented which extends the *relative motion* (REMO) GKD approach developed to identify motion patterns in groups of MPOs (Laube and Imfeld, 2002; Laube et al., 2004; Laube et al., 2005). The REMO approach allows the user to search large volumes of data for instances of pre-defined patterns, which are constructed on the basis of existing knowledge about the motion of the objects under study. However, the user has no means by which to estimate the significance (for example, the uniqueness) of the extracted patterns. The number and the extent of patterns found may depend significantly on both the motion data and the parameterisation of the pattern detection process. Many more patterns may be identified in a space where motion is constrained, for example on a football pitch, than in the seemingly chaotic motion of children in a playground. A central question for GKD is therefore, how can we assess the significance of patterns extracted from such cases?

Our approach focuses on the use of Monte-Carlo simulations to generate synthetic lifelines constrained by the statistical properties of real observational data. Pattern occurrence in the simulated data is then compared with observational data to assess the potential significance of the patterns. Finally, having identified potentially interesting patterns we return to the observational data to investigate whether these patterns have meaning in terms of the system under investigation.

The paper is structured as follows. Section 2 provides a literature overview on geographic knowledge discovery in general, the REMO GKD approach, pattern interestingness, and on different potential approaches to simulating geospatial lifelines. In Section 3 the central ideas of our evaluation approach are introduced. The data used in the study are introduced in Section 4. Section 5 describes the methodology for producing constrained random walks to produce synthetic lifelines, and their use in Monte-Carlo simulation for numerical experiments. The results of these experiments are presented in Section 6, and their meaning and application to the observational data is discussed in Section 7 before examining the general applicability of these results to the field of GKD.

2 Background

2.1 Geographic Knowledge Discovery

Knowledge discovery in databases (KDD) is a set of methods for identifying high-level knowledge from low-level data in the context of large datasets. As an interdisciplinary approach it integrates methods from machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for experts systems, data visualisation, and high-performance computing (Fayyad et al., 1996). KDD moves beyond the traditional domain of statistics to accommodate data normally not amenable to statistical analysis. Descriptive statistics usually involve small and relatively noiseless numeric data scientifically sampled from a large population with a specific question in mind. In contrast, KDD is designed for data collected and stored in many scientific or enterprise databases which are potentially noisy, non-numeric, and incomplete.

The KDD process is interactive and iterative, involving data selection, data cleaning and pre-processing, data reduction and projection, exploratory analysis and data mining, and result interpretation and evaluation. The central belief of KDD is that information is hidden in very large databases in the form of interesting patterns (Miller & Han, 2001). Data mining, that is the application of specific algorithms for extracting patterns from data, is thus just one component of the overall KDD process, (Fayyad et al., 1996).

Miller and Han (2001) identified unique needs and challenges for integrating KDD into GIScience. They argue that geographic data has unique properties that require special KDD and specifically data mining approaches. As examples of uniquely spatial properties they list the geographic measurement framework (Euclidian geometry and topology), spatial dependency and heterogeneity, the complexity of spatio-temporal objects and relationships, and diversity of involved data types. Hence, the need for specific geographic knowledge discovery and geographic data mining techniques is proposed as an important area for research.

2.2 The REMO approach

REMO GKD seeks to discover motion patterns in the lifelines created by a group of MPOs. Suitable lifeline data consists of the trajectories of a set of MPOs each featuring a list of fixes (x,y,t). The approach is based on two key features (Figure 1):

- a transformation of the lifeline data (a) to an analysis matrix (b,c) featuring motion attributes (i.e. speed, change of speed or motion azimuth),
- followed by matching of formalized patterns on the matrix (d).

The geospatial lifelines in Figure 1 (a) represent the tracks of four GPS-collared deer. Deer O_1, moves with a constant motion azimuth of 45° during an interval from t_2 to t_5, i.e. four discrete time steps of length ∂t, and shows a *constancy* pattern. The four deer with the same motion azimuth of 45° at the time t_4 show a *concurrence* pattern. An MPO anticipating the motion of others shows a *trend-setter* pattern. Thus, a *trend-setter* pattern is constructed by searching for *constancy* in conjunction with *concurrence* (e.g. deer O_1 anticipates at t_2 the motion azimuth 45° that is reproduced by all other MPOs at time t_4).

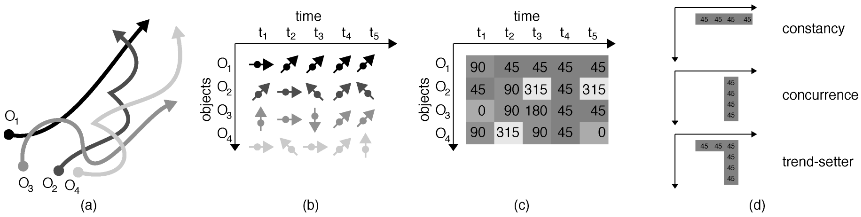


Fig. 1. **The REMO GKD approach**. The geospatial lifelines of four MPOs (a) are used to derive at regular intervals the motion azimuth (b). In the REMO analysis matrix consisting of classified motion attribute values (c) generic motion patterns such as *constancy*, *concurrence* or *trend-setter* are matched (d).

2.3 Pattern Interestingness

Finding interesting patterns in large databases requires that pattern discovery algorithms are developed, where the patterns found are considered to be an anomalous feature of the data, departing from what might be expected (Hand, 2004). The perennial problem of pattern discovery is to define what an anomalous pattern is within the context of a very large dataset. A feature of discovered patterns is considered by the KDD community to be their *interestingness* (Silberschatz & Tuzhilin, 1996, Padmanabhan, 2004) – that is to say do the patterns have some meaning to the user within the context of the KDD process. Within KDD criteria for interesting patterns are divided into those which are objective and subjective. Objective measures depend solely on the structure of the pattern and the underlying data. Subjective measures depend also on the class of users exploring the data, bearing in mind that patterns that are of interest for one user class, may be of no interest to another class (Silberschatz & Tuzhilin, 1996). Tuzhilin & Silberschatz (1996) identify two reasons why a pattern is interesting from a subjective point of view: *unexpectedness,* which indicates how surprising the pattern is to a user, and *actionability* which indicates whether the user can act on the pattern to his/her advantage. However, in very large datasets a pattern may be unexpected, but nonetheless not unusual. A key need in examining unexpectedness is therefore to quantify how often a supposedly unexpected pattern occurs within the dataset, and better, the likelihood of such an unexpected pattern occurring. In this paper we use interestingness to assess the potential significance of mined patterns in a GKD process.

2.4 Simulating Motion

Although the potential of large datasets containing many thousands of MPOs for use in GKD is clear, the actual availability of very large datasets has been limited in the past for two main reasons. Firstly, these datasets were simply not available for a combination of legal (e.g. privacy) and technical reasons (Pfoser & Theodoridis, 2003), and secondly research in KDD and GKD has proceeded on the basis of the expected increase in volume of data. Thus, considerable work has been done in generating synthetic motion data in a diverse range of subjects including ecology, transport modelling and database research. Importantly, the use of synthetic data allows the generation of arbitrary spatial and temporal granularities. A summary of key elements of this work is introduced below.

In population ecology synthetic motion data is widely used to study the distribution and abundance of organisms. See Turchin (1998) for an introduction to the measuring and modelling population redistribution in animals and plants. In transport modelling synthetic motion data emerges from traffic simulation for transportation planning (Nagel et al., 2000). In the spatio-temporal data base management (STDBMS) community synthetic data has been used generating semantics-based trajectories of MPOs for designing novel data types and access methods for spatio-temporal databases (Pfoser & Theodoridis, 2003) and for benchmarking of proposed database models (Brinkhoff, 2002; Saglio & Moreira, 2001).

A variety of different approaches exist to the simulation of motion data. These range from relatively simple random walk procedures, commonly used in ecology (Turchin, 1998) through to complex multi-agent simulations (Batty et al. 2003). Whatever the approach, there are two key features: firstly, the simulation of the MPOs themselves and secondly, the environment within which they can act, interact and react.

In our case, the environment has no influence on the MPOs other than constraining their movement within some bounding box. Our datasets consist solely of fixes, with no associated environmental attributes for a given time and place. Thus, we focus here on the representation of MPOs as simple point objects performing some kind of *random walk*. Within the random walk framework MPOs link moves, each featuring a duration, a speed and a direction. A move can be influenced by the direction of its previous move, local conditions or attraction towards some destination. Typically (but not necessarily), there is a stochastic element involved in selecting the next move (Turchin, 1998). For *correlated random walks* the direction of the current move does affect the direction of the next move. Step length and turning angle of the subsequent move are drawn from some stochastic frequency distribution, for instance in the case of direction with turning angles concentrated around zero. A *biased random walk* includes the influence of an absolute direction, i.e. a long distance attraction. The work presented in this paper adopted a *constrained random walk*, where constraints from the analysis of real observation data are used to set turning angle distributions and step sizes (Wentz et al., 2003).

Monte Carlo simulation is widely used in GIScience and other fields to quantify the effects of variables with some uncertainty (Heuvelink, 1999). In combination with constrained random walks and the associated frequency distributions, Monte Carlo techniques offer the possibility of generating multiple lifelines for use in experiments with synthetic data and comparison with observational data.

2.5 Aims of this paper

From this background we can summarise the key aims of this paper.

- How can we estimate the interestingness of motion patterns mined by the REMO GKD approach?
- Is the use of Monte Carlo simulated populations of synthetic motion data an appropriate strategy for evaluating a motion pattern detection approach?
- Do interesting patterns as defined through these techniques correspond with events which have meaning to domain experts?

3 Evaluating the REMO approach

Hand (2004) defined pattern discovery as the search for anomalous features of the data, departing from the expected. However, the challenge lies in defining the expected. If an unexpected pattern is also unique, then the user can reasonably examine its interrelationships and infer meaningfulness. If, on the other hand, an unexpected pattern occurs many times how can we first assess if it is not only unexpected, but unusual given the parameterisation of the data? The first step in assessing such unexpectedness is therefore to develop techniques to simulate lifelines which have similar statistical properties to the observed data. Our approach to simulating lifelines uses constraints derived from the objects under study to derive synthetic motion data of arbitrary spatial and temporal granularities.

Silberschatz and Tuzhilin (1996) propose unexpectedness as a measure of interestingness of patterns. They argue that patterns are interesting because they contradict our expectations, given by our system of beliefs. One option to capture the beliefs is to formulate a statistical hypothesis. The degree of belief is then defined as a significance level for which the statistic for a certain test is on the borderline of acceptance of the hypothesis. The main drawback of this statistical approach to indicate unexpectedness is that not any belief can be formulated as a testable hypothesis (Silberschatz & Tuzhilin, 1996). For this reason we replace the statistical hypothesis with Monte Carlo simulations, choosing the detour of numerical experiments to learn about “the expected” from the stochastic properties of the simulations.

Since our aim is to assess the interestingness of REMO patterns, we hypothesise that interesting patterns will have a higher ratio of occurrence to those created in simulated CRWs. Monte Carlo simulations are used to generate a population of simulated CRWs, which can be compared with the observational data. Our underlying assumption is that those patterns which appear to be outliers from the stochastic properties of the simulations are those which we can attach some initial meaning to, prior to further investigation by the user.

The technique used to generate lifelines is crucial. Synthetic lifelines created by a complete random walk would not be suitable in assessing the interestingness of motion patterns since within the complete random walk framework every step is completely independent of the previous step with respect to both direction and step length. Such lifelines do not represent the characteristic motion of either migrating animals or the Swiss districts. Behavioural ecologists have addressed this problems by using simulation techniques that consider the characteristics of MPOs and their motion behaviour (Bergmann et al., 2000; Byers, 2001; Turchin, 1998; Wentz et al., 2003).

For many simulation problems the use of stochastic frequency distributions known from the literature which characterise the objects and their motion are a suitable choice. However, if such general knowledge is not available, one possible solution is to derive frequency distributions directly from the observation data at hand. For example, Wentz et al. (2003) used two different techniques to create synthetic trajectories filling gaps in the incomplete lifelines of primate species for analysing home range and daily ranging patterns. They illustrated that the use of constrained random walk produced reasonable approximations of field observations and for certain species performed significantly better in analysis than simple linear interpolation between the known points in the trajectory.

4 Data

Two contrasting datasets are used in evaluating our approach. One of these consists of wildlife data, with a relatively small number of objects, but well understood behaviour. The other dataset are a classic example of spatialisation, where aspatial attributes are projected into a geographic space. This second dataset consists of many more objects and time steps than our wildlife data, and is a typical example of a dataset where GKD might be expected to reveal hitherto unseen patterns.

4.1 Porcupine Caribou

The Porcupine Caribou Herd Satellite Collar Project is a cooperative project that uses satellite radio collars to document seasonal range and migration patterns of the Porcupine Caribou Herd (PCH) in northern Yukon, Alaska and Northwest Territories (NWT). Details about the Porcupine Caribou Project including the original tracking data can be found under http://www.taiga.net/satellite.

This data set has been selected because it has the ideal granularity suitable to analyse seasonal migration. Furthermore seasonal migration is a known behaviour within the PCH and well documented (Fancy et al., 1989; Fancy & Whitten, 1999; Griffith et al., 2002). A key test of the usefulness of the REMO system must be its ability to identify known patterns, such as this migration. Just as with any other wildlife field study, this observation data set does not provide a complete coverage of all individuals over the whole study period. Animals die, loose their GPS collar, or had their collars removed. Thus, our experiments focus on a subgroup of the herd, consisting of 10 individuals simultaneously tracked over almost two years, starting from March 2003 (see Figure 2a).

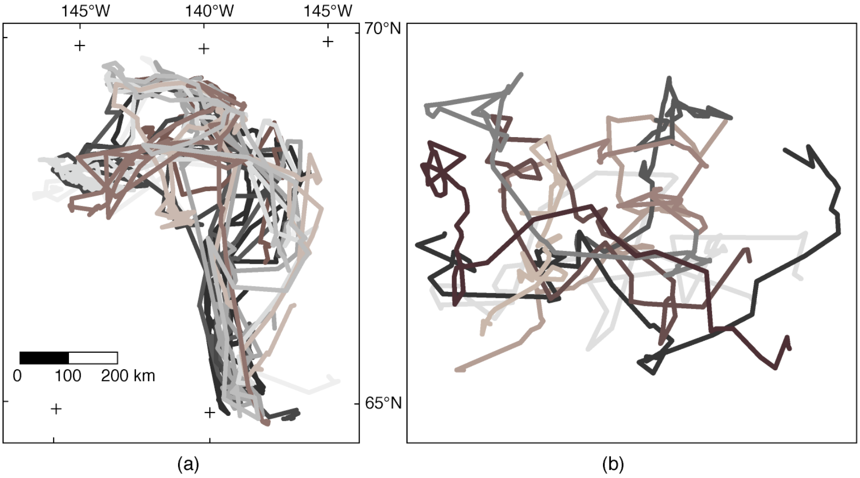


Fig.2. **Observed and simulated Porcupine Caribou data.** (a) Tracks of 10 individuals of PCH between March 2003 and December 2004. (b) One sample of r = 100 simulated populations of constrained random walks. The simulated tracks feature the same frequency distributions for step length and turning angle as the observed tracks in (a). Please note that the observed paths overlap considerably and thus appear much more densely packed than the simulated tracks starting at random positions in space.

4.2 Abstract Data Points

Frequent popular referendums in Switzerland (approx. 8-10 per year) allow researchers to make detailed inferences about value conflicts within society. Hermann and Leuthold (2001) developed an inductive approach to explore basic ideological conflicts in Switzerland. By performing factor analysis on referendum data at the district level of all 158 federal referendums held between 1981 and 1999, they hypothesised a structure of mentality, which was interpreted as being composed of three dimensions: political left vs. political right, liberal vs. conservative and ecological vs. technocratic (Hermann and Leuthold, 2003). In these two dimensional ideological spaces the 185 districts can be positioned at intervals of one year, from 1981 until 1999 and their movement through this hypothesised ideological space over a period of 20 years followed (see Figure 3a).

This data set has been chosen for two reasons. Firstly it features a large number of simultaneously tracked MPOs (n = 185). Secondly, the districts of a Canton (member states of the Swiss Federation) show a certain institutional and cultural similarity, potentially expressed in similar motion behaviour within the abstract space, – potentially mirrored in detectable patterns. The major limitation of this data set lies in its small number of time steps (t = 20).

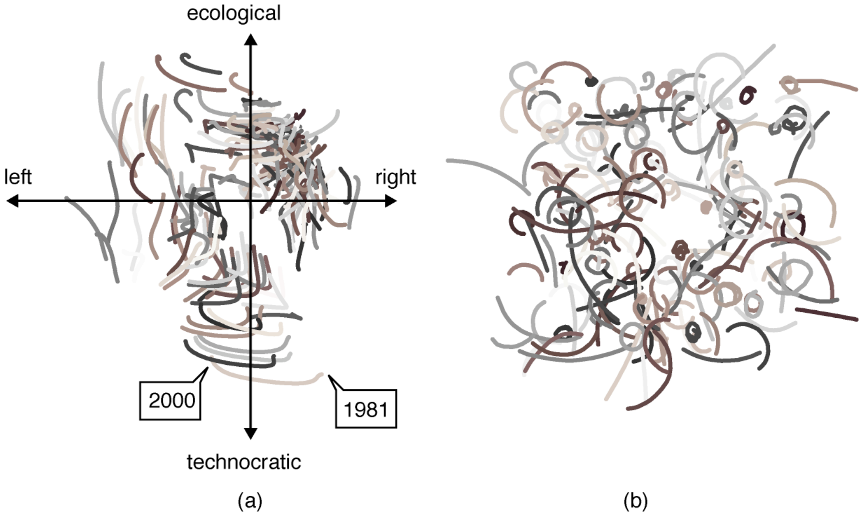


Fig. 3. **Observed and simulated Swiss District data.** (a) Tracks of 185 Swiss political districts moving from 1981 to 2000 between the ideological poles left vs. right and ecological vs. technocratic. (b) One sample of r=100 simulated populations of constrained random walks. As in figure 2 also in this plot the simulated tracks start at random locations in space. Note that both simulated and observed tracks, keep a preferred turning direction (see for example the subset of 12 districts in Figure 7), expressing curly shapes.

5 Methodology

5.1 Simulating Lifelines using Constrained Random Walk

To generate simulations with properties as close as possible to the observed motion phenomena, we use constraints derived from the observation data. The constraints are given as frequency distributions of step length and direction change (Figures 4-7). This establishes an empirical link between the simulated and real data, without which the utility of the random walk model is considered as being severely restricted (Turchin, 1998).

For the construction of synthetic trajectories using a constrained random walk, we generate for every synthetic MPO its own step length and direction change array, for instance of size d = 1000, corresponding to the observed frequency distributions of the MPO. The construction of any new point P_{t+1} of the trajectory requires selection of a sample s_{t+1} from the step length distribution as well as a sample a_{t+1} from the direction change angle distribution, where the sample is selected at random from the distribution (see Fig. 8).

A crucial element of constrained random walks are the initial conditions. In both case studies we use random starting locations for the artificial MPOs since the algorithms of the REMO GKD approach evaluated in this paper only consider relative, and not absolute, positions of MPOs. Differing approaches were used for the starting direction. For the PCH data random starting angles were applied since a long temporal interval, with observed distributions lying between –180 and +180 degrees exists (Fig. 7). For the moving districts, by contrast, the actual first step directions of the observation data were used to initialise the random walk model. This option was chosen because the starting angle has a large influence on the outcome the simulation under the given conditions of a short temporal interval and a narrow direction change distribution (Fig. 6).

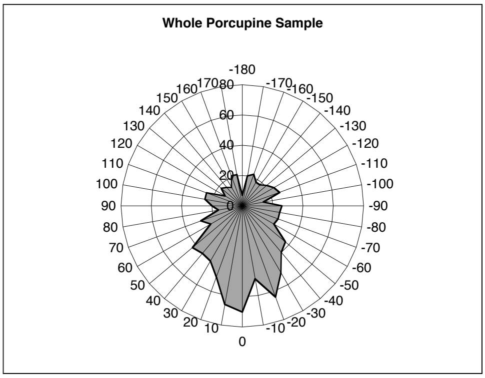


Fig. 4. **Direction change frequency distribution of PCH sample.** The radar plot illustrates an overall directional persistence for moving straight on (0°).

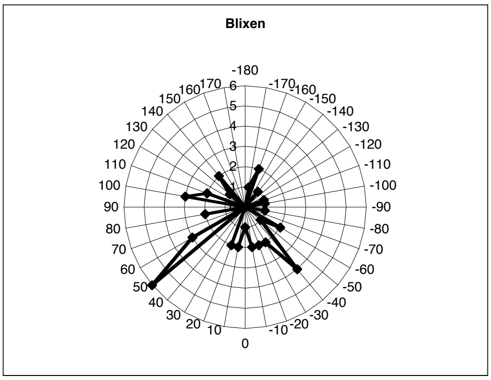


Fig. 5. **Direction change frequency distribution of caribou individual Blixen.** Individual direction change distributions can considerably diverge from the overall frequency distribution in Figure 4.

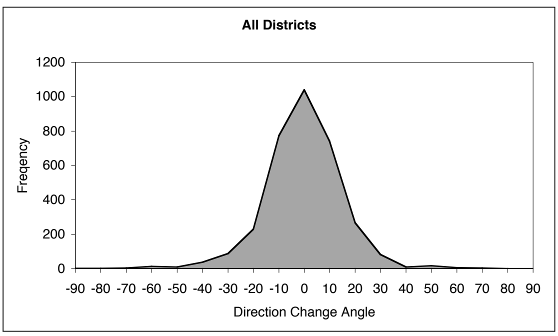


Fig. 6. **Direction change frequency distribution of total Swiss political districts sample.** The direction changes of the moving districts are much more concentrated around 0 (moving straight on), and are thus not plotted in a radar plot.

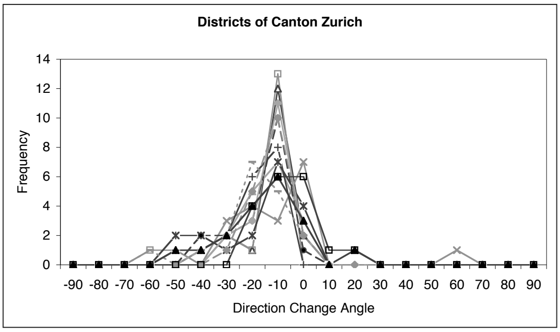


Fig. 7. **Direction change frequency distribution of districts of Canton Zurich.** This subset of 12 districts illustrates the diversity in the direction change distribution among individual MPOs.

5.2 Monte-Carlo Experiments

The design of the MC experiment is twofold: In a first step we generate a set of *r* data sets with synthetic lifelines of *n* MPOs (see Figure 9). In a second step we run the REMO pattern matching algorithms on the generated lifeline data systematically varying the properties of the pattern matching process (see Fig. 10).

Generating synthetic data. For a direct comparison of the observed data with the simulated trajectories we reproduced 100 synthetic groups as described above (see Figures 2b and 3b).

- **PCH.** n = 10 MPOs, t = 96 weeks, ∂t = 2 weeks, r = 100 simulation runs.
- **Moving districts.** n = 185 MPOs, t = 20 years, ∂t = 1 year, r = 100 simulation runs.

REMO GKD with variable configurations. We performed REMO pattern detection experiments for the patterns *constancy, concurrence*, and *trendsetter* with respect to motion *azimuth* and *speed* for both case studies. Three nested loops control the experiments (see Fig 10).

- In the innermost loop we vary the pattern extension *p*. In the case of the pattern over time (e.g. constancy) *p* expresses the pattern length, that is, over how many time steps a pattern exists. *p* takes values from 2 time steps (i.e. 4 weeks with the PCH data, 2 years with the Swiss district data) to a maximum number representing the whole time period (t = 96 weeks, t = 20 years). In the case of patterns across objects (e.g. concurrence), *p* stands for the number of involved MPOs, ranging from 2 to the total number of MPOs in the group (10 caribou, 185 districts respectively).
- This procedure we apply *r* times for all the *r* synthetic motion data sets in the next exterior loop. Thus *r* = 100 represents the number of Monte Carlo runs.

- Finally, in the outermost loop, we repeat the MC experiment for the predefined set of attribute classifications c. For the experiments involving motion azimuth we used $c = [2, 4, 8, 16, 36, 360]$ attribute classes (i.e. two classes with 180° intervals, four classes with 90° intervals, eight classes with 45° intervals, …). For the experiments involving speed we used $c = [2,4,8,16]$ classes.

5.3 Assessing interestingness of REMO patterns

For every pattern matching run of the configuration (c,r,p) both the number of possible and the number of found patterns are calculated. The ratio of found to possible patterns is then used to provide a standardised means to indicate how many patterns were found per pattern matching run of configuration (c,r,p).

The maximum number of patterns that can be packed into a matrix is dependent on the number of attribute classes and the pattern extent. This maximum number decreases as pattern extent increases. The influence of the number of classes is less obvious, since it depends on the pattern type. For constancy the number of classes has no influence on the maximum number of possible patterns since the matrix could be packed with an alternation of attributes with only two classes. For concurrence the maximum number of patterns per column corresponds to the number of attribute classes, since ordering of the column is irrelevant in instantiating a pattern

All result plots have the following structure (see for example Figure 12). The mean of the ratio of the number of patterns found for the Monte-Carlo simulations to the possible number of matches for this pattern length are plotted as whisker plots. The x-axis represents the extent p of the patterns, referred to as pattern *length* for patterns over time (e.g. constancy) and pattern *width* for pattern across objects (e.g. concurrence). The ratio on the y-axis represents the number of patterns found. The whisker plots for the simulation data ($n = 100$) feature mean, 25th and 75th percentile, as well as minimal and maximal values.

Following Tuzhilin and Silberschatz (1996) we define that patterns are worthy of investigation when we find more matches in the observed data than in the simulated data. We further suggest, that when the ratio of the number of patterns observed is an order of magnitude different from the simulated data some qualitative notion of interestingness can be extracted. For example, in Figure 12 (bottom) the ratio of found/possible has a value of about 0.2 for the observed data, and a mean of around 0.02 for the simulated data with a pattern width of 5. We would suggest therefore that concurrence patterns of width 5 with 8 azimuth classes are worthy of investigation in this (PCH) dataset.

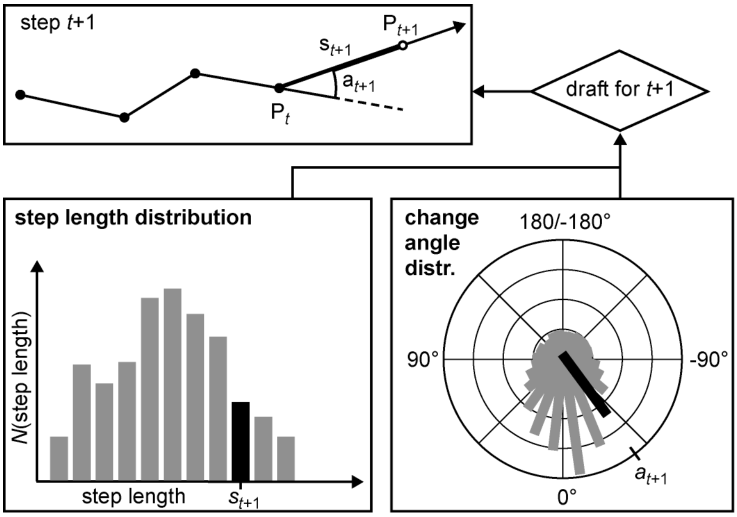


Fig. 8. **The next step.** For the construction of any step at $t+1$ two samples from the constraining distributions are needed. First, a step length s_{t+1} and second a direction change angle a_{t+1}

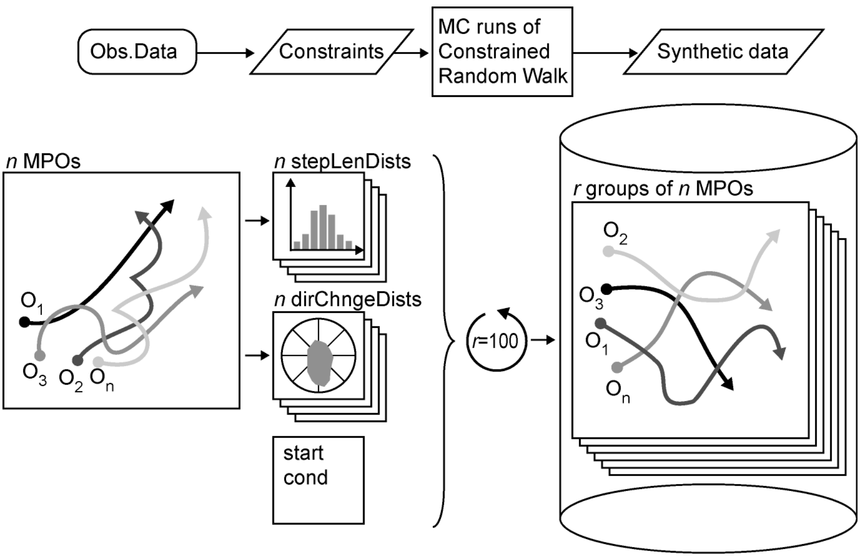


Fig. 9. **Monte Carlo simulations of constrained random walks.** The constraints' *step length distribution* (stepLenDist) and *direction change distribution* (dirChngeDist) for the random walk are first derived from an observation data set consisting of the trajectories of n MPOs. Each of the r = 100 Monte Carlo runs generates thereafter a synthetic data set of the constrained random walks of n MPOs.

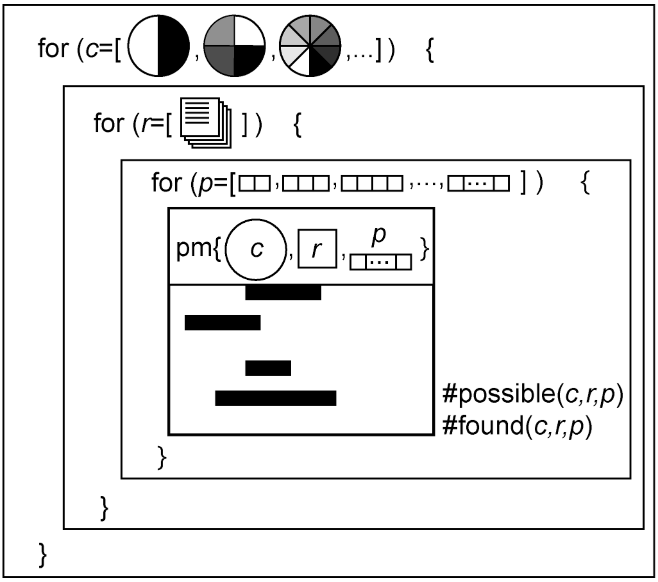


Fig. 10. **Numerical experiments.** The numerical experiments repeat the pattern matching process (pm) for all *r* synthetic data sets systematically varying the pattern extent *p* and the attribute classification *c*. For every pattern matching run the number of found patterns and the number of possible patterns is stored for subsequent statistical analysis.

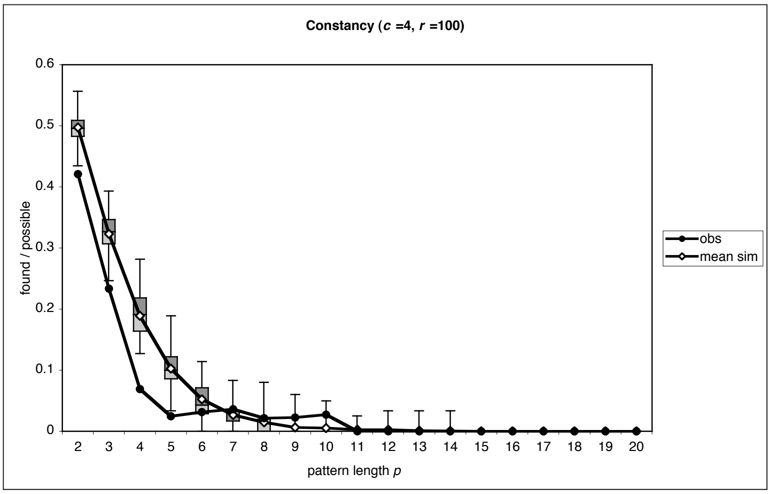
6 Results of the numerical experiments

In this paper we present results from experiments performed, with varying pattern extension and attribute granularity, in order to evaluate the interestingness of the following patterns: *constancy*, *concurrence* and *trendsetter* (Table 1)

Fig.	Data	Motion property	pattern	attribute granularities	Pattern extent *p*
11	Porcupine Caribou Herd	azimuth	constancy	4 (top), 8 (bottom)	length 2-20 time steps
12	Porcupine Caribou Herd	azimuth	concurrence	4 (top), 8 (bottom)	width 2-10 individuals
13	Porcupine Caribou Herd	azimuth	trendsetter	4 (top), 8 (bottom)	3 and 4 steps anticipation, width 2-10 individuals
14	Porcupine Caribou Herd	speed	constancy	8	length 2-20 time steps
15	Porcupine Caribou Herd	speed	concurrence	8	width 2-10 individuals
16	Swiss political districts	azimuth	concurrence	4 (top), 8 (bottom)	width 2-65 districts

Table 1. Configurations of the numerical experiments

The results we present in this section involve the motion properties *motion azimuth* and *speed*, focussing on azimuth. Although having performed experiments with a wide range of different attribute granularities, we selected for this paper the two most commonly chosen granularities for indicating directions, that is four and eight cardinal directions (corresponding to the geographical directions N, E, S, and W, as well as N, NE, E, SE, S, SW, W, NW, respectively).



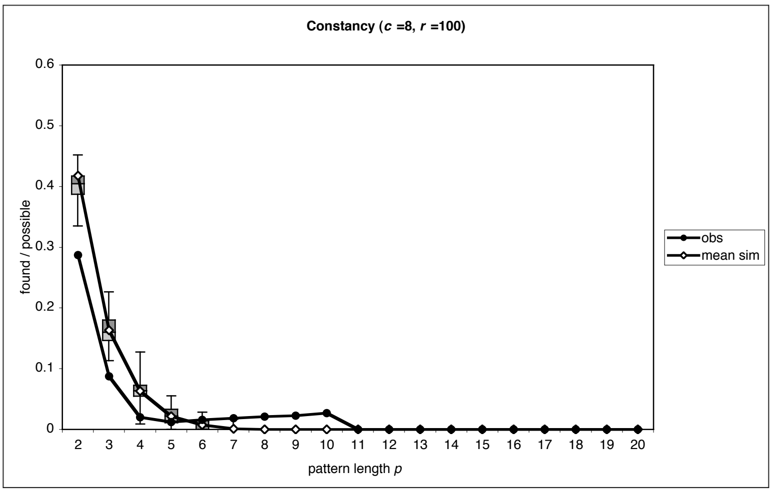
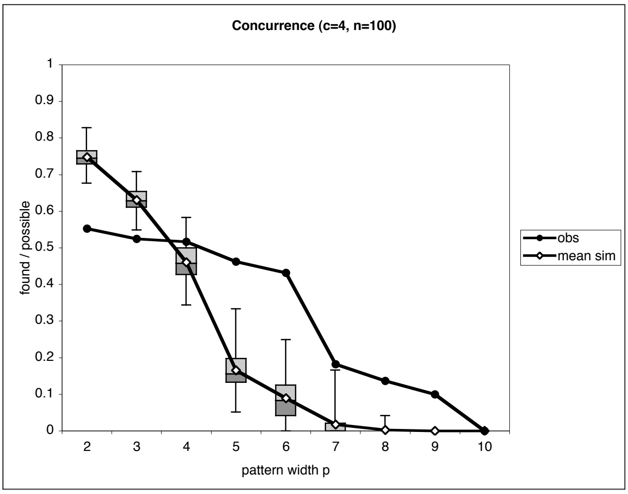


Fig. 11. **Detecting motion azimuth constancy patterns in PCH.** The success of the pattern matching process is very similar for the observed and the simulated data, rapidly declining with increasing pattern length *p*.

A simple relationship between granularity and pattern length is present in all plots – in both observed and simulated data – finer attribute granularities have less simulated (and also less observed) patterns as pattern the extent increases. See for example Figure 12 to compare the number of *concurrence* patterns found with azimuth granularity of four and eight classes respectively. A further feature of changing granularity is that as granularity increases, in our data, we find more observed than simulated patterns as the pattern extent increases. For example, in Figure 12 (top), with 4 classes, more short patterns are present in the simulated data than the observed.



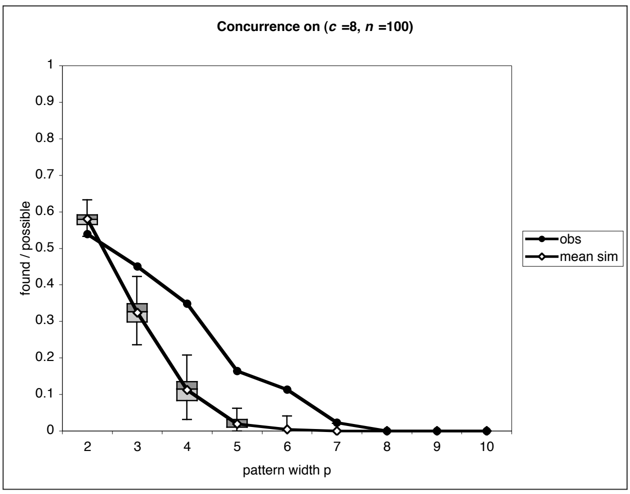
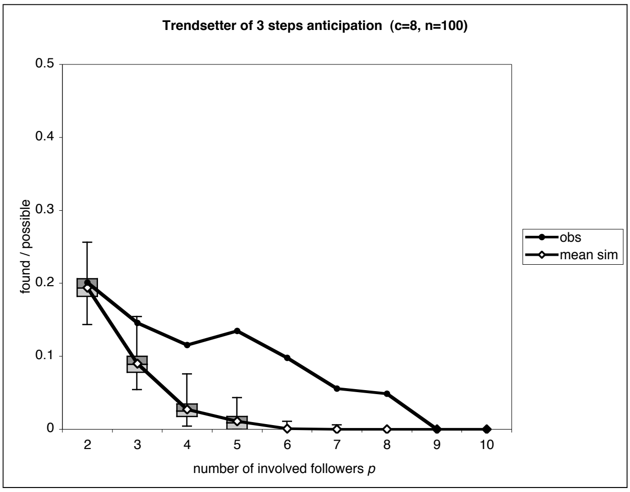


Fig. 12. **Detecting motion azimuth concurrence patterns in PCH.** The obvious deviation of the found/possible ratio for the observed and the simulated data allows to assign interestingness to patterns. The greater the difference between the ratios for observed and simulated, the higher the assigned interestingness.

Due to the discrete nature of the REMO patterns as well as the analysis matrix, the number of maximally possible patterns per matrix does not decrease smoothly with increasing patterns extents. In contrast, the number of possible patterns per matrix expresses discrete jumps, reflected as artefacts in the result plots (see for example Figure 16 top). For example, with 4 azimuth classes and 185 districts, maximally 4 instances of a concurrence of width 46 are possible at a single time step (4*46=184<185). Increasing the concurrence width by one to 47, all of a sudden only 3 instances of concurrence can be packed in a time step (4*47=188>185). However, since simultaneously also the number of found patterns is expected to decrease for the same reason, the ratio *found/possible* remains a reliable measure to indicate the success of a pattern.



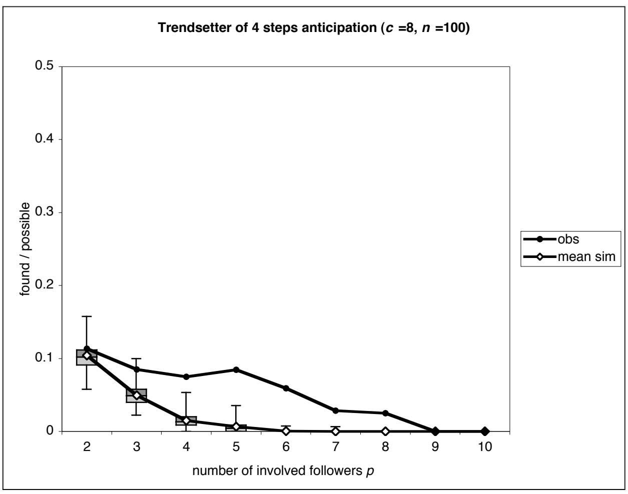


Fig. 13. **Detecting motion azimuth trend-setter patterns in PCH.** Both plots are based on c = 8 motion attribute classes, the difference lies in the length of the required anticipation of the trendsetter, 3 and 4 time steps, respectively. Generally, the found/possible ratio of trend-setter lies below the results for constancy and concurrence, for both observed and simulated. This is not surprising, keeping in mind that the requirements for complex patterns such as trendsetter are much higher. However, observed lies considerably above simulated, allowing to assign interstingness to found trendsetter patterns.

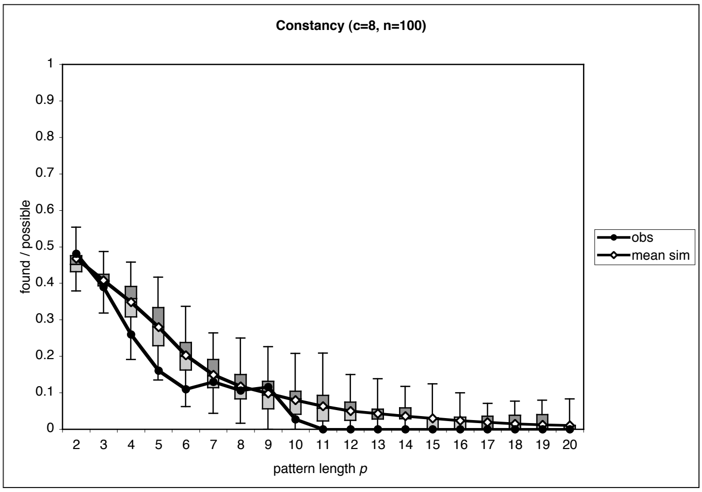


Fig. 14. **Detecting speed constancy patterns in PCH.** The variance of the ratio found/possible for the simulated data is much larger than in most plots referring to motion azimuth. However, the observed value lies in the variance range and does therefore not allow to assign interestingness to found speed patterns.

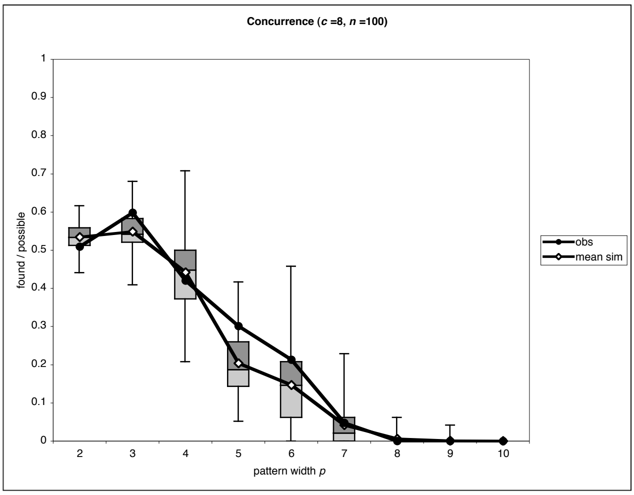
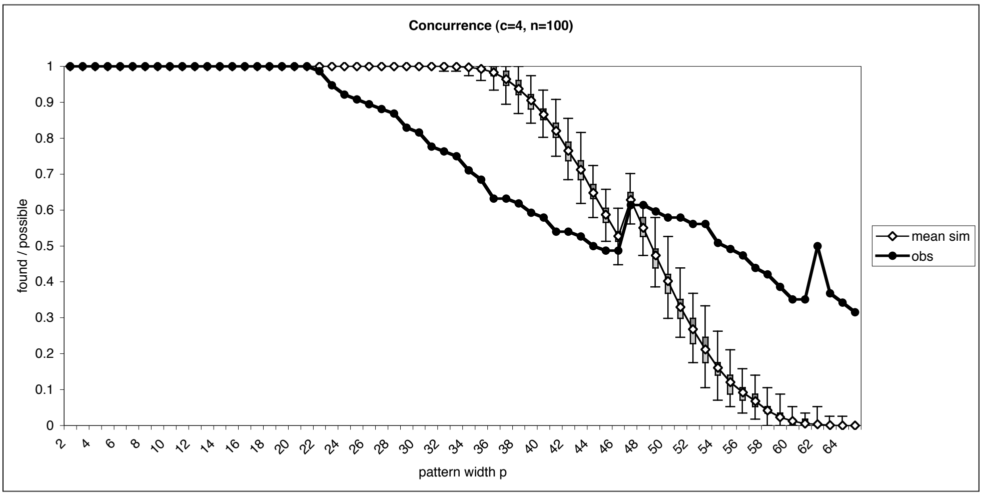


Fig. 15. **Detecting speed concurrence patterns in PCH.** Just like constancy, also concurrence does not show considerable deviation between the observed and the simulated values.



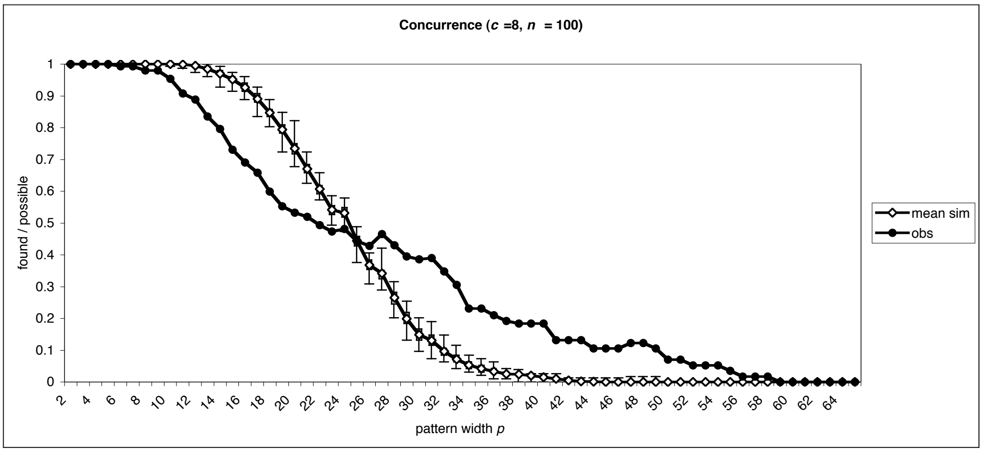


Fig. 16. **Detecting motion direction concurrence in the Swiss political districts.** For both attribute granularities significantly more patterns are found than expected from the simulation, after a threshold given by the crossing of both graphs. The sharp jump in both graphs is an artifact caused by the computation of the ratio found/possible.

7 Discussion

In discussing the results shown in Section 6, we consider two different aspects of the study. We firstly seek to interpret the plots derived from comparison of the properties of the observed data with the simulations, without consideration of the context of the data. That is to say, we do not take note of whether our MPOs are caribou or Swiss political districts in this first discussion. Secondly, the results from the simulations are used to identify potentially interesting patterns within our data. We look for these patterns and then consider them in the light of the context of the data – for example, the spring migration of caribou, or the socio-political shifts in the society in the case of our political districts. Thirdly, we discuss the proposed method in the light of our findings.

7.1 Discussion of features of plots

Consider first the case of constancy illustrated in Figure 11 (top), with 4 classes, where we find more short patterns in the simulated data than the observed. This simple result leads us to conclude that we can attach no meaning to these short patterns in the observed data. By contrast, in Figure 12 (bottom), with 8 classes, almost all pattern width have more matches in the observed data than the simulated. Following our definition of interestingness these patterns are unexpected and therefore interesting.

Figure 11 shows an example where simulation shows that constancy patterns in the observed PCH data are not worthy of further investigation. Here, it is clear that for a granularity of 4 classes the number of simulated patterns is greater than or equal to the number of observed patterns. For patterns derived from 8 attribute classes, the situation is slightly less clear cut. A small ratio of long patterns occur in the observed data, but since the simulated data still suggest these are unusual (no long patterns being found in the simulated data), they may merit further examination.

The numerical experiments for concurrence affirm a straightforward hypothesis about the relation between attribute classification and pattern extent (pattern length/width). The number of MPOs divided by the number of attribute classes gives a lower threshold of pattern interestingness. With 185 MPOs and 4 classes (8 classes), concurrence patterns with a width of less than 46 (below 23) are expected simply by randomly sampling motion azimuths from the 4 classes (8 classes).

It is at a first glimpse rather surprising that below the threshold many less concurrence patterns are found in the real data that would be expected from the experiments (see Fig. 12 (top), Fig. 16). Assuming that the observed motion is indeed coordinated one can easily explain the lack of real patterns below the threshold. Where coordinated motion does in fact exist, then wider concurrence patterns will be found. This in turn reduces the number of occurrences of narrower patterns, as seen in the observed data in Figures 12 (top) and 16. By contrast, in the simulated data where the angular distribution is evenly distributed across motion azimuth classes, narrow patterns are more likely to be found.

7.2 Analysis of results with respect to case studies

7.2.1 Porcupine Caribou Herd (PCH)

Figure 17 depicts the REMO matrix for the PCH trajectories with *c*=8 colour coded motion azimuth classes. The pattern matching results illustrated in Figures 18 und 19 are performed on this matrix.

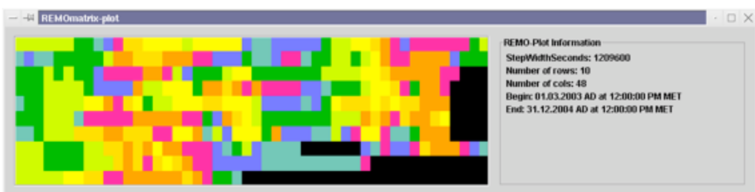


Fig. 17. **REMO matrix for PCH motion azimuth.** Rows represent caribou individuals, columns represent time steps ∂t two weeks. The c=8 azimuth classes are colour coded, using a circular colour ramp ranging from green (N), blue (E), orange (S), to yellow (W).

Fancy and Whitten (1989) refer to two short directed seasonal migration patterns in the PCH seasonal motion, a spring migration in May and a fall migration in September. The PCH "commutes" between calving areas near the Beaufort Sea in early summer and the winter areas in the South. Having in mind that the sampling rate of ∂t = 2 weeks is rather coarse compared to the duration of the whole seasonal migration, caribou moving straight on for longer temporal intervals (∂t = 4, 6, 8 weeks) cannot be expected in their trajectories.

From Figure 12 we derived a concurrence width of five as being a meaningful pattern. Thus, the results illustrated in Figure 18 showing concurrence patterns of at least width five can be assigned interestingness. All patterns appear in the migration seasons, where the PCH heads Northwest (spring; light and dark green colours) and Southwest (late summer and early fall; orange, red colours), respectively.

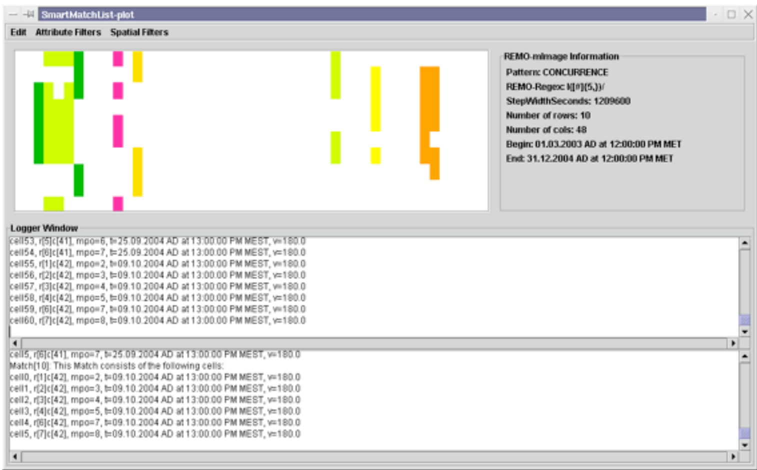


Fig. 18. **Concurrence matches in PCH.** This pattern detection session searched for concurrence patterns including at least 5 caribou. Such patterns could be found in both northward spring migration and southward fall migration.

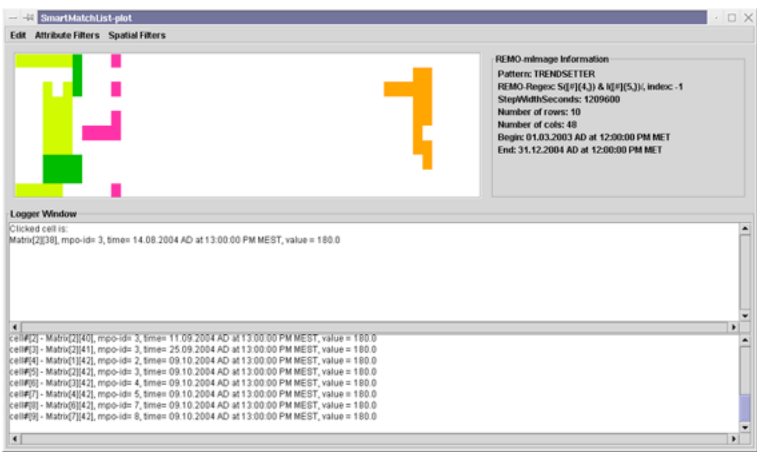


Fig. 19. **Trend-setter matches in PCH.** Trend-setting caribou individuals anticipating 3 time steps in advance the seasonal migration of at least 5 followers.

Bearing in mind that very long periods of straight motion are not expected, we investigated whether we could find trend-setting caribou, anticipating the seasonal migration before the rest of the herd. Indeed, as shown in Figure 19 (top), we find trend-setting individuals with at least five followers anticipating the spring and fall migration. For instance, the caribou Lynetta, anticipated in mid August 2004 the fall southward migration of more than 5 followers caribou in early September (see Figure 19, orange pattern in right part of plot). Again, we have chosen the number of followers that expressed the largest difference between observed and simulated in Figure 13, here p = 5 individuals.

Investigating the REMO matrix representing c = 8 speed classes, one would expect to find many concurrence patterns both in the fast migration season (blue) as well as in the slower sedentary seasons (red) (Figure 20). As shown in Figure 15 the pattern matching process results indeed in relatively high ratios of found/possible. However, since the variance of the ratio in the simulated data is rather high and the ratio found in the observation data is in the same range, we can't assign interestingness to the found concurrence patterns. Also with constancy (Figure 14) the observed patterns lie in the rather wide range of the simulation. Again, with a coarse temporal sampling rate and many different activity periods in the caribou year (two migration seasons, scattering in the calving season, mid-August dispersal (Fancy, et al. 1989), we cannot expect to find significant speed constancy patterns.

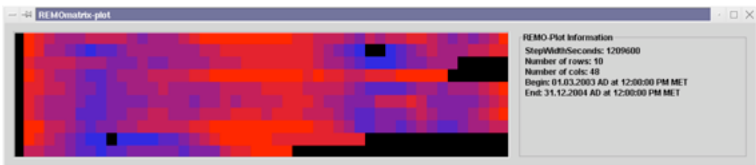


Fig. 20. **REMO matrix for PCH speed.** As a general overview the speed matrix illustrates two fast (blue) migration periods each season in 2003 and 2004, respectively, with intermitted slower sedentary periods.

7.2.2 Swiss political districts

Investigating the motion direction of the Swiss political districts in Figure 16, we see that meaningful concurrence patterns appear to occur after a threshold length of 48 for four classes and for a threshold of 25 with eight classes, where significantly more patterns than predicted by the simulations are observed. Laube et al. (2005) suggested a concurrence of 45 districts as a hint for a political left-right divergence, ideologically separating the German and the Latin part of Switzerland in the 1980s and 1990s. Now, having numerical evaluation experiments at hand, we can indeed attach interestingness to that pattern (Figures 21 and 22). Investigating the Figure 16 bottom, we see that concurrence patterns with a width of 45 appear in the simulated pattern with a ratio of found/possible of almost 0, in the observed data with around 0.1 respectively, hence satisfying the required order of magnitude. In the same publication Laube et al. (2005) identified a concurrence pattern of width 18. Considering the results in Figure 16 we would not assign any interestingness to that pattern anymore and rather expect to find patterns with this extent purely by chance.

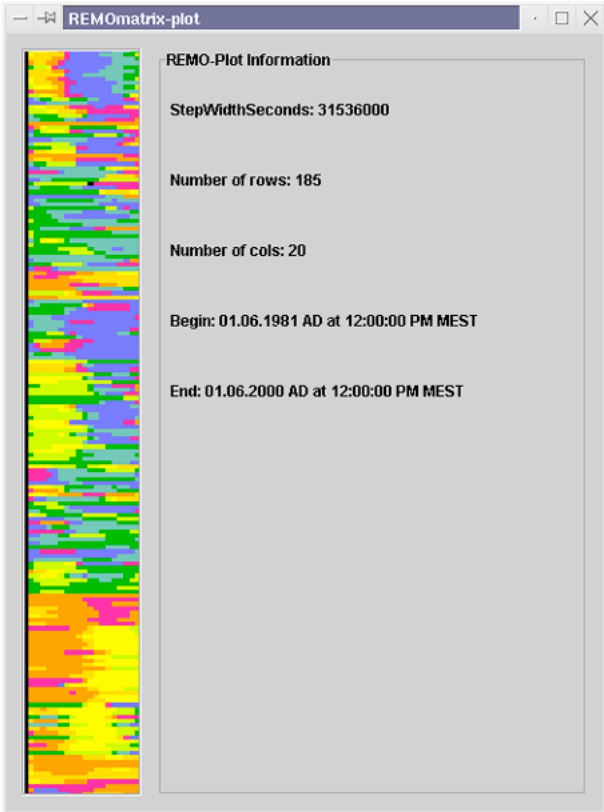


Fig. 21. **REMO matrix for Swiss political districts motion azimuth.** Rows represent 185 districts, columns represent 20 years from 1981 until 2000. The c = 8 azimuth classes are colour coded, using a circular colour ramp ranging from green (N), blue (E), orange (S), to yellow (W).

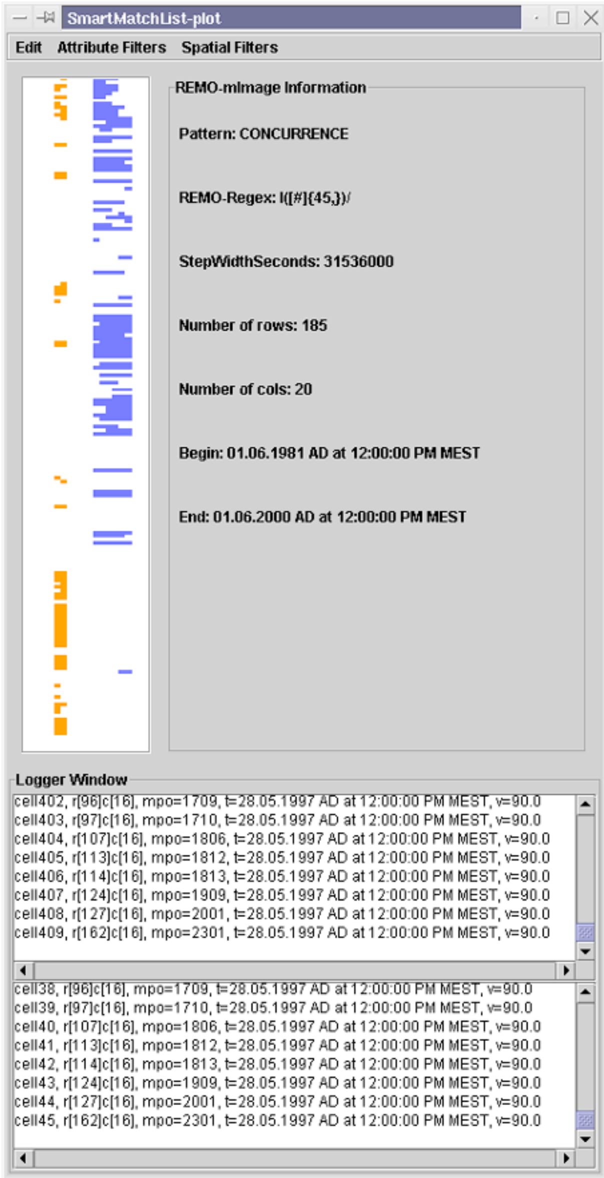


Fig. 22. **Concurrence matches in Swiss political districts motion azimuth.** This pattern detection session searched for concurrence patterns including at least 45 concurrently moving districts. Two sequences of several concurrence patterns could be found during this pattern detection session.

7.3 Discussion of proposed method

One might first argue that using observation data to parameterise the simulation and thereafter comparing the simulation with actual observation data has an inherent danger, in that any such reasoning is circular. However, our aim is not to analyse the statistical properties of the static trajectories of MPOs, but rather to consider the unexpectedness of emerging motion patterns. Comparison of our data with a random sample would increase the likelihood of a pattern being unexpected (since any spatial or temporal autocorrelation in motion azimuth and step length would be removed) and unexpected patterns would be more likely. Thus, arguably, the use of correlated random walks based on the distribution of observational data is a stricter test of unexpectedness.

The selection of constraints for the numerical experiments is crucial. In our work, we have used all available MPOs to define constraints. The initial direction of movement was defined differently in the two datasets. For the caribou, where turning angles span ±180° the initial direction of movement was randomly assigned and had no influence on the results due to the large number of time steps. However, in the case of the Swiss political districts the use of a random starting angle, coupled with the narrow distribution of turning angles and the small number of timesteps resulting in motion with similar characteristics in terms of speed, but widely varying motion azimuths. By using observed motion azimuth this influence was reduced. A further step in measuring the potential influence of initial azimuth would be to randomly subdivide the data to derive both constraints and initial directions of movement for comparison with the remaining observed data.

As shown above, small samples of lifelines may represent very specific motion characteristics, such as in the case of the moving districts. For instance, if we simply observe the caribou during migration periods, the constraints given by the frequency distribution would describe for most MPOs a rather rapid and linear motion behaviour. Thus, in such data, motion azimuth concurrence could be expected and such patterns would be of no interest. In contrast, a sedentary period with randomly foraging MPOs would simply not express any pattern to assess interestingness. However, by using lifeline populations from constraints featuring both migratory and sedentary periods, we are able to identify instances of such patterns if they are actually present in the observational data. It will be important to revisit the methodology as larger samples of observational data, with more normal distributions become available.

Two datasets are used in this study. In the case of the caribou, only ten MPOs are present. In the case of the Swiss political districts a total of 185 MPOs exist, for a relatively short number (20) of timesteps. Initial research (Laube and Imfeld, 2002; Laube et al., 2005) sought to identify patterns in these datasets through visual inspection of patterns. However, though these datasets appear initially small, when the number of potential combinations of pattern, granularity, and motion type are considered very large volumes of data are created. Furthermore, Kwan (2000) argues convincingly that the user's ability to identify meaningful patterns is quickly swamped using visual exploration. Therefore, we here seek to marry techniques based on ideas for dealing with large datasets with commonly used visualisation approaches. The use of these techniques shows considerable promise, allowing patterns to be identified which were not apparent only through inspection, but will require further testing as large datasets become available.

8 Conclusions

8.1 Main contributions

In this paper we presented a method to assess qualitative data mining measures of interestingness in geographic knowledge discovery. The GKD approach under study is the REMO GKD, developed to detect motion patterns in the lifelines of moving point objects. We propose a procedure to estimate the interestingness of the motion patterns found in lifeline observation data. Therefore, we first generate populations of constrained random walks using Monte Carlo simulations. Secondly, we compare the success of the pattern matching process on observation data with that in the simulated data.

Previous publications discussing the REMO GKD approach suggested that there is a strong relation between the parameterisation of the pattern detection process and the number of patterns found (Laube & Imfeld, 2002; Laube et al., 2005). Having the numerical experiments of this work at hand, we have now evidence to reinforce this hypothesis.

We suggest that such an approach – testing the interestingness of patterns by performing numerical experiments to measure their unexpectedness is one which can be transferred to other techniques which seek to extract information from very large datasets. Furthermore, as was shown by our experiments, we would argue that caution must be attached to patterns that are presumed to be interesting only through visual inspection.

8.2 Implications for studies of moving point objects

The MC experiments with both the Caribou and the Swiss district data point out that performing the REMO GKD process a lower pattern length/width threshold must be adopted depending on the classification schema.

The evaluation method introduced in this work can be used to examine useful configurations of pattern matching sessions (i.e. attribute granularities, pattern lengths), which may not be obvious. The method helps, for example, in selecting suitable compromises between granularity and information.

In general, methods similar to our approach, can be used in (geographic) knowledge discovery to focus the search for meaningful patterns. Hence, not only the (putatively biased) knowledge of the potential user may be used to configure the GKD process, but also evidence from numerical experiments.

Motion pattern studies like REMO are crucial to better understand why ‘spatial (and spatio-temporal) is special’ with respect to knowledge discovery, – to learn more about the ‘G’ in GKD. Carried out carefully, GKD has a huge potential for applications in spatialising socio-economic processes or wildlife studies.

8.3 Outlook

In future work we will investigate more complex patterns, increase the number of MC runs, and include more wildlife and sports data in our numerical experiments. Furthermore we intend to explore alternative random walk approaches to optimally approximate the MC simulations to the observed motion process. For example, in the case of football players biased correlated random walk models (goals as attraction points) and transition matrices of Markov Chains may be evaluated (Jones and Smith, 2001).

Acknowledgements

The authors wish to thank Michael Hermann and Heiri Leuthold, University of Zurich, for providing the Swiss districts data, and the Porcupine Caribou Technical Committee, the Porcupine Caribou Management Board and the Wildlife Management Advisory Council (North Slope) for providing the excellent caribou data. The authors would furthermore like to acknowledge invaluable inputs from Nikos Koutsias, Robert Weibel, Stephan Imfeld, University of Zurich.

References

Batty, M., & Jiang, B. (1999). Multi-agent simulation: new approaches to exploring space-time dynamics within GIS. Technical report, Centre for Advanced Spatial Analysis, University College London.

Batty, M., Desyllas, J., and Duxbury, E. (2003). The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades. *International Journal of Geographical Information Science*, *17*(7), 673–697.

Bergman, C.M., Schaefer, J. A., & Luttich, S. N. (2000). Caribou movement as a correlated random walk. *Oecologia, 123*(3), 364-374.

Brinkhoff, T. (2002). A Framework for Generating Network-Based Moving Objects, *GeoInformatica, 6*(2), 153-180.

Burstedde, C., Klauck, K., Schadschneider, A. & Zittartz, J. (2001). Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A, 295*(3-4), 507-525.

Byers, J.A. (2001). Correlated random walk equations of animal dispersal resolved by simulation. *Ecology, 82*(6), 1680-1690.

Dijkstra, J., Jessurun, A.J., &Timmermans, H.J.P. (2000). A Multi-Agent Cellular Automata System for Visualising Simulated Pedestrian Activity. In S. Bandini, and T. Worsch (Eds.), *Theoretical and Practical Issues on Cellular Automata, Proceedings of the Fourth International Conference on Cellular Automata for Research and Industry* (pp. 29-36). Berlin-Heidelberg: Springer.

Dykes, J.A. & Mountain, D.M. (2003). Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and application. *Computational Statistics and Data Analysis, 43*(4), 581-603.

Fancy, S.G. & Whitten, K.R. (1991). Selection of calving sites by Porcupine herd caribou. *Canadian Journal of Zoology, 69*, 1736-1743.

Fancy, S.G., Pank, L.F., Whitten, K.R. & Regelin, W.L. (1989). Seasonal movement of caribou in arctic Alaska as determined by satellite. *Canadian Journal of Zoology, 67*, 644-650.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54.

Frank, A. U. (2001). Socio-Economic Units: Their Life and Motion. In Frank, A.U., Raper, J. & Cheylan, J.P. (Eds.), *Life and motion of socio-economic units, GISDATA 8* (pp. 21-34). London: Taylor & Francis.

Griffith, B., Douglas, D.C., Walsh, N.E., Yound, D.D., McCabe, T.R., Russell, D.E., White, R.G., Cameron, R.D. & Whitten, K.R. (2002). The Porcupine caribou herd. In Douglas, D.C., Reynolds, P.E. & Rhode, E.B. (Eds.), *Arctic Refuge coastal plain terrestrial wildlife research summaries* (pp. 8-37), U. S. Geological Survey, Biological Resources Division, Biological Science Report USGS/BRD/BSR-2002-0001.

Griffiths, G.H. & Mather, P.M. (2000). Remote Sensing and landscape ecology: landscape patterns and landscape change. *Int. J. Remote Sensing, 21*(13-14), 2537-2539.

Hand, D. J. (2004). Pattern Discovery. *Journal of Applied Statistics, 31*(8), 883-884

Hermann, M., & Leuthold, H., (2001). Weltanschauungen und ihre soziale Basis im Spiegel der Volksabstimmungen. *Swiss Political Science Review, 7*, 39–63.

Hermann, M., & Leuthold, H., (2003). *Atlas der politischen Landschaften. Ein weltanschauliches Portrait der Schweiz.* Zurich: vdf-Verlag.

Heuvelink, G. B. M. (1999). Propagation of error in spatial data modelling with GIS. In Longley, P., Goodchild, M. F., Maguire, D. J. & Rhind, D. W. (Eds.), Geographical Information Systems (pp. 207-217). New York, NY: John Wiley & Sons.

Hornsby, K., & Egenhofer, M. J. (2002). Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence, 36*(1-2), 177-194.

Iwase, S. & Saito, H. (2003). Tracking Soccer Players Based on Homography among Multiple Views. In Ebrahimi, T & Sikora, T. (Eds.), *Visual Communications and Image Processing* (pp. 288-292), Proceedings of SPIE.

Jones, P. W., & Smith, P. (2001). *Stochastic Processes: An introduction.* London, UK: Arnold Publishers.

Kwan, M.-P. (2000). Interactive geovisualization of activity-travel patterns using three dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C, 8*(1-6), 185-203.

Laube, P., & Imfeld, S. (2002). Analyzing relative motion within groups of trackable moving point objects. In Egenhofer, M. J. & Mark, D. M. (Eds.), *Geographic Information Science, Second International Conference, GIScience 2002, Boulder, CO, USA, September 2002*, *LNCS 2478* (pp. 132–144). Berlin-Heidelberg: Springer.

Laube, P., Imfeld, S., & Weibel, R. (2005). Discovering Relative Motion Patterns in Groups of Moving Point Objects. *International Journal of Geographical Information Science*. *19*(6), 639–668.

Laube, P., Van Kreveld, M., & Imfeld, S. (2004). Finding REMO - detecting relative motion patterns in geospatial lifelines. In Fisher, P. F. (Ed.), *Developments in Spatial Data Handling, Proceedings of the 11th International Symposium on Spatial Data Handling* (pp. 201-214), Berlin-Heidelberg: Springer.

Mark, D. M. (1998). Geospatial Lifelines. *Dagstuhl Seminars No. 98471, Integrating Spatial and Temporal Databases.*

Mark, D. M. (2003). Geographic Information Science: Defining the Field. In Duckham, M., Goodchild, M. F. & Worboys, M.F. (Eds.), *Foundations of Geographic Information Science* (pp. 3-18). London: Taylor and Francis.

Miller, H. J. & Han, J. (2001). Geographic data mining and knowledge discovery: An overview. In Miller, H. J. & Han, J., (Eds.), *Geographic data mining and knowledge discovery* (pp. 3–32). London: Taylor and Francis.

Miller, H.J. (2003). What about people in geographic information science?, *Computers, Environment and Urban Systems*, *27*(5), 447–453.

Nagel, K., Esser, J., & Rickert, M. (2000). Large-scale traffic simulations for transportation planning. In Stauffer, D. (Ed.) *Annual Reviews of Computational Physics VII* (pp. 151-202), World Scientific Publishing Company.

Padmanabhan , B. (2004). The Interestingness Paradox in Pattern Discovery. *Journal of Applied Statistics, 31*(8), 1019-1035.

Pfoser, D., & Theodoridis, Y. (2003). Generating semantics-based trajectories of moving objects. *Computers, Environment and Urban Systems, 27*(3), 243–263.

Saglio, J.-M., & Moreira, J. (2001). Oporto: A realistic Scenario Generator for Moving Objects, *GeoInformatica, 5*(1), 71-93.

Silberschatz, A., & Tuzhilin, A. (1996). What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 970-974.

Turchin, P. (1998). *Quantitative Analysis of Movement: Measuring and Modelling Population Redistribution in Animals and Plants*, Sunderland, Massachusetts: Sinauer Publishers.

Wentz, E. A., Campell, A. F., & Houston, R. (2003). A comparison of two methods to create tracks of moving objects: linear weighted distance and constrained random walk. *International Journal of Geographical Information Science, 17*(7), 623–645.

Wood, J. (2002). *Java Programming for Spatial Sciences*. London: Taylor and Francis.